Improved Bayesian methods for detecting recombination and rate heterogeneity in DNA sequence alignments

Alexander V. Mantzaris

Doctor of Philosophy

Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh

2011

Abstract

DNA sequence alignments are usually not homogeneous. Mosaic structures may result as a consequence of recombination or rate heterogeneity. Interspecific recombination, in which DNA subsequences are transferred between different (typically viral or bacterial) strains may result in a change of the topology of the underlying phylogenetic tree. Rate heterogeneity corresponds to a change of the nucleotide substitution rate. Various methods for simultaneously detecting recombination and rate heterogeneity in DNA sequence alignments have recently been proposed, based on complex probabilistic models that combine phylogenetic trees with factorial hidden Markov models or multiple changepoint processes. The objective of my thesis is to identify potential shortcomings of these models and explore ways of how to improve them.

One shortcoming that I have identified is related to an approximation made in various recently proposed Bayesian models. The Bayesian paradigm requires the solution of an integral over the space of parameters. To render this integration analytically tractable, these models assume that the vectors of branch lengths of the phylogenetic tree are independent among sites. While this approximation reduces the computational complexity considerably, I show that it leads to the systematic prediction of spurious topology changes in the Felsenstein zone, that is, the area in the branch lengths configuration space where maximum parsimony consistently infers the wrong topology due to long-branch attraction. I demonstrate these failures by using two Bayesian hypothesis tests, based on an inter- and an intra-model approach to estimating the marginal likelihood. I then propose a revised model that addresses these shortcomings, and demonstrate its improved performance on a set of synthetic DNA sequence alignments systematically generated around the Felsenstein zone.

The core model explored in my thesis is a phylogenetic factorial hidden Markov model (FHMM) for detecting two types of mosaic structures in DNA sequence alignments, related to recombination and rate heterogeneity. The focus of my work is on improving the modelling of the latter aspect. Earlier research efforts by other authors have modelled different degrees of rate heterogeneity with separate hidden states of the FHMM. Their work fails to appreciate the intrinsic difference between two types of rate heterogeneity: long-range regional effects, which are potentially related to differences in the selective pressure, and the short-term

periodic patterns within the codons, which merely capture the signature of the genetic code.

I have improved these earlier phylogenetic FHMMs in two respects. Firstly, by sampling the rate vector from the posterior distribution with RJMCMC I have made the modelling of regional rate heterogeneity more flexible, and I infer the number of different degrees of divergence directly from the DNA sequence alignment, thereby dispensing with the need to arbitrarily select this quantity in advance. Secondly, I explicitly model within-codon rate heterogeneity via a separate rate modification vector. In this way, the within-codon effect of rate heterogeneity is imposed on the model a priori, which facilitates the learning of the biologically more interesting effect of regional rate heterogeneity a posteriori. I have carried out simulations on synthetic DNA sequence alignments, which have borne out my conjecture. The existing model, which does not explicitly include the within-codon rate variation, has to model both effects with the same modelling mechanism. As expected, it was found to fail to disentangle these two effects. On the contrary, I have found that my new model clearly separates within-codon rate variation from regional rate heterogeneity, resulting in more accurate predictions.

Acknowledgements

I would like to first thank my supervisor Dirk Husmeier for his time and dedication that he has put into supervising my research. His energy and enthusiasm has been a crucial element for development. Next I would like to thank my family and especially my father for finding interest in listening to updates on the status of the project. I would like to thank another supervisor as well, Frank Wright, for inspiring many interesting thoughts in the discussions we have had. Finally I would like to thank everyone in the ANC institute and BIOSS.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Alexander V. Mantzaris*)

Table of Contents

Chapter 1

Introduction and methodology

Organisms which reproduce asexually like bacteria, amoebas, and viruses have different evolutionary patterns than organism which undergo sexual recombination. There is a linear trace of the genetic material from one generation to the next under asexual reproduction. Having more than one parent in reproduction results in more than one path in back tracking the genetic history of an organism.

Recently studies of bacteria and viruses have shown that genes can be horizontally transferred between cells as shown in Robertson *et al.* (1995). This process of evolutionary changes is also referred to as recombination. In latter sections more information will be given to explain this effect. For the moment it suffices to say that this can allow more diverse and abrupt changes to the evolutionary progress of the organism.

1.1 Molecular Sequence Data

In all organisms the genetic material is stored and encoded in either DNA (Deoxyribonucleic acid) or RNA (Ribonucleic acid). RNA is used as the main source of storage of genetic material for viruses. DNA is used by higher level organisms. There are 4 bases for the encoding of genetic material in DNA. These are Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). These base molecules are specific in that G and C bond together, and A and T bond together as pairs. In RNA the difference is that Thymine does not exist and is replaced with Uracil (U) as a base pair in those positions. Mutations between sequences of DNA or RNA are due to substitutions which are considered to be stochastic.

In the single nucleotide mutations, which are called point mutations, there

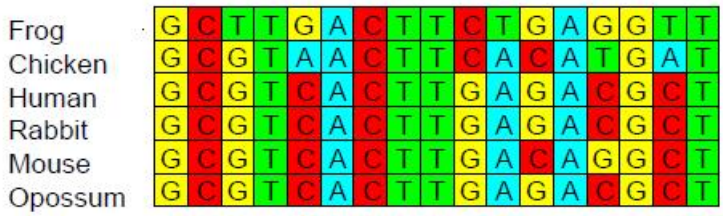


Figure 1.1: DNA sequence alignment

The image shows a subset of a sequence alignment of the haemoglobin gene. The 6 species listed on the left of the alignment name the DNA stretches which run horizontally. Taken from *http://www.bioss.ac.uk/staff/dirk/talks/lectureDougArm0203.pdf*.

are two types of substitutions. Nucleotides A and G are in a group of molecules called purines and nucleotides C and T in a group called pyrimidines. Mutations between nucleotides within the groups are called transitions, and mutations between the groups are called transversions. Making this differentiation is useful since the frequencies of *transitions* and *transversions* are not the same.

In comparing different DNA sequences, analogous pairs of nucleotides are considered in an alignment. The alignment of different length sequences is a separate topic and is introduced in Durbin *et al.* (1998). Figure 1.1 shows a short section of a sequence alignment for a set of different species. The different nucleotide sequences are arranged by species in rows. The columns of the nucleotides sequence alignments are used to compare the differences between the genomes of each species.

1.2 Phylogenetic trees

Phylogenetics has the purpose of reconstructing the evolutionary relationships between organisms from a sequence alignment. Many sources of evidence for an evolutionary reconstruction can be used such as fossils, phenotypic traits, and others. This thesis is concerned with relationships built upon molecular data. The goal is to use the sequence alignments to construct a tree whose structure depicts the evolutionary relationship between the sequences. Both the evolutionary history and the ancestry can be seen from a phylogenetic tree. In building phylogenetic trees we assume that the force of evolution is fundamentally

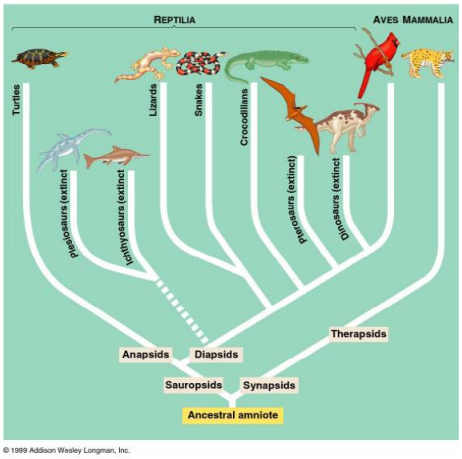


Figure 1.2: Example phylogenetic tree

A basic example of a phylogenetic tree used to demonstrate the basic features and information it represents is shown here. What can be seen are the bifurcation points which are instances of speciation where the lineage continues with time to 2 different species. This common point can be traced back to the time when the 2 species share a common ancestor. This can be useful for tracing the age of new traits seen after a speciation event. The distance between nodes represents time which is flowing upwards. The leaves of the tree are referred to as *taxa*, and the *extant* species are the leaves which can be observed (here the dinosaurs are extinct). Taken from 1999 Addison Wesley Longman, Inc.

probabilistic when causing the mutations to emerge, and this probabilistic model or a suitable approximation is desired when building the tree. The term, *taxa* is used for the organisms whose genetic material is present in the alignment. Figure 1.2 shows an example of a phylogenetic tree where the extant species are shown with simple diagrams.

Phylogenetic trees are bifurcating, binary, tree structured graphical models. The bifurcation points are nodes which represent common ancestral genomes. These ancestral nodes bring together the pair of species which have deviated further down the tree with time. The leaves of the tree represent the *extant* species which can be observed (unless extinct). Bifurcations are important, because the time passed since the separation between new features can be useful information. The branch lengths are the distance between nodes on the graphical model (representation) of the phylogenetic tree and are used to denote *phylogenetic time*. This shows the number of expected mutations per site, and is the product of the

rate of mutation and time.

There are different possible arrangements of the bifurcations of the phylogenetic tree. Different arrangements result in different clustering of species' distances between them. Figure 1.3 shows the graphic model representation of phylogenetic trees estimated from a sequence alignment. This sequence alignment subset is depicted in figure 1.1. A chosen arrangement is termed *topology*. Different topologies may place bifurcations in a different order, and this can be interpreted as grouping species to be closer or further apart between each other. For example in one topology a human may be inferred to be closely related to a rabbit, and in another topology the presence of an additional bifurcation in the tree between humans and the human-rabbit common ancestor would suggest that another species was more closely related to humans than rabbits. Subfigure a) shows a *rooted* tree whereas subfigure b) shows an *unrooted* tree. Rooted trees are directed according to time. The root depicts the beginning of time and the distance from it (the time passed since the time of the root). Unrooted trees have no root for a starting time point. The number of topologies which are possible is derived from the number of sequences, or taxa, that are being compared. This is because of the permutations of the bifurcations for the taxa. As the number of sequences increases the number of possible tree topologies increases super-exponentially. For m DNA sequences there are

$$(2m-3)!! \tag{1.1}$$

rooted topologies for the taxa, and

$$(2m-5)!! \tag{1.2}$$

different unrooted topologies. The double factorial denoted is similar to standard factorial $m! = m(m-1)(m-2)\dots(2)(1)$ in that the difference in the values between products is 2 rather than 1, $m!! = m(m-2)(m-4)\dots(3)(1)$. To denote a topology the S variable is used that can take on one of the topologies of the set $1,\dots,(2m-5)!!$. These relationships are derived in the appendix A.1.

Previously in this section rate heterogeneity and branch lengths were mentioned. Both of these have related effects on the phylogenetic tree. Rate heterogeneity (without considering the branch lengths yet) scales the size of the tree to be larger or smaller and is done uniformly. The ratio of the individual lengths between nodes is kept the same and there is a coefficient scaling the size of all of

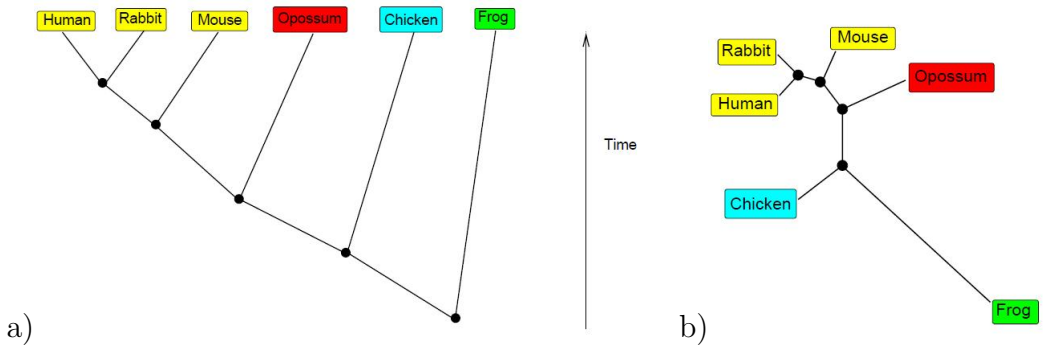


Figure 1.3: Example of rooted and unrooted phylogenetic trees as graphical models
An example sequence alignment is shown in figure 1.1 and is used to generate the phylogenetic trees in the two subfigures a) and b). From the alignment we can see that the first species, the frog, has smallest number of common nucleotides with any other sequence and as a result it is drawn further away from any other species. The human and the rabbit have the fewest number of mutations between them and therefore have the smallest distance between them and their common ancestor. These lengths between nodes are called branch lengths and will be spoken about in more depth in later sections. Subfigure a) shows a rooted tree that has an estimated root from the time point where the divergence and speciations began for all the extant species. Subfigure b) shows an unrooted tree connecting the taxa. A common ancestor for all the taxa is not inferred, and neither is a time point for the initial starting point of the evolution process. Figures taken from talks of Dirk Husmeier.

those lengths. This coefficient is referred to as a rate factor and is denoted by ρ. A ratefactor value less than 1 is indicative of negative selection pressure, a value close to 1 for neutral selection and a value greater than 1 for positive selection.

The individual branch lengths are denoted with, w_i, and the vector of all the branch lengths in the phylogenetic tree is denoted by $\mathbf{w}$. For an unrooted tree of m taxa (from m sequences), there are $m-2$ internal nodes and $2m-3$ branch lengths, so $1 \leq i \leq (2m-3)$. The mutations observed for a particular topology are the number of mutations between species separated by a bifurcation. These mutations are converted into a branch length representation of the *phylogenetic time* that has passed (this is discussed in more detail in subsections 1.4, 1.5).

1.3 Models of nucleotide substitution

A distribution for the substitution of nucleotides $x \in A,C,G,T$ into another nucleotide or into itself is required for a probabilistic model of phylogenetics. $P(*|x,w)$, is the distribution for any mutation that is conditional on the present nucleotide and the branch length which is the *phylogenetic time*. Phylogenetic time is the product of the rate of mutation and time, and gives an expectation for the number of mutations. Figure 1.4 shows in subfigure a) the process of substitutions based on the present nucleotide, and b) a possible graphical demonstration of the probability of the substitution process along phylogenetic time. We can see that when no phylogenetic time has passed the probability of the present nucleotide to be found there is 1. At infinite phylogenetic time we can see the rate of change of the substitution probability becoming zero as the memory of the present state has an effect that converges to zero in the infinite approximation. From the graph it can be seen that all the nucleotides have equal probability at the largest value on the horizontal axis which is the branch length. The branch length is the product of time with the mutation rate which gives an intuitive use for the expected number of mutations to be seen. From the graph we can see that there is a faster increase for the probability of nucleotide A to be substituted into G which is a transition as introduced in subsection 1.1. A transversion in this case would be a substitution into a nucleotide C or T. In general, transitions are a more probable substitution than a transversion. This will be elaborated in subsection 1.3.3.

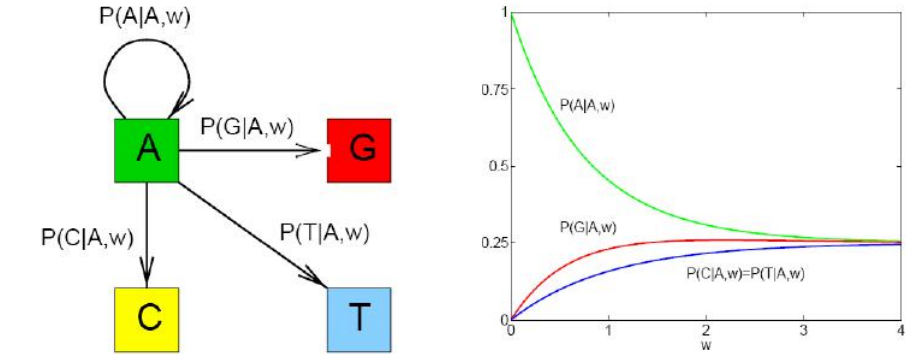


Figure 1.4: Substitution Model

An example of how a probabilistic approach considers nucleotide substitutions. Subfigure a) shows the model of a substitution process with the Markov independence, that the probability of a certain nucleotide being substituted with any other nucleotide is dependent on the present nucleotide. The branch length is used as a parameter, in the nucleotide substitution model, to change the distribution of the substitutions. Subfigure b) shows the probabilities of the substitutions as they vary with the branch length, w. The branch length represents the length of the expected substitutions and so is proportional to time and the mutation rate, $w = (\rho) \times (time)$. On the vertical axis we have the probability and the horizontal axis is the branch length which is *phylogenetic time*. When no time has passed there is no probability to find a different substituted nucleotide, and at infinite phylogenetic time we see no rate of change in the nucleotides' substitution probabilities, as the stationary distribution has been achieved. Figures taken from Husmeier *et al.* (2005a).

1.3.1 Nucleotide substitution models

There are various nucleotide substitution models for modelling mutations in phylogenetic trees. In this section the details of nucleotide substitution models are given, these form the framework for the following work. The subsection 1.5 and figure 1.4 describe the concepts of the nucleotide substitution models and its relation to likelihood methods in phylogenetics.

In this section we let $y_i(t) \in A,C,G,T$, stand for the nucleotide at a site i and at a time t. For convenience and where necessary to aid the reader, the use of some of these symbols may differ in later sections. The index of the site in the alignment of length N takes on values from 1 till N, $i \in 1,\ldots,N$. The theory of *homogeneous Markov chains* underlies the assumptions used to build the model of nucleotide substitutions.

1. The process is Markov:

   $$P(y_i(t+\Delta t)|y_i(t), y_i(t-\Delta t),\ldots) = P(y_i(t+\Delta t)|y_i(t))$$

2. The Markov process is homogeneous in time:

   $$P(y_i(s+t)|y_i(s)) = P(y_i(t)|y_i(0))$$

3. The Markov process is the same for all positions: $P(y_i(t)|y_i(0)) = P(y_k(t)|y_k(0)) \forall i,k \in 1,\ldots,N$

4. Substitutions at different positions are independent of each other: $P(y_1(t),\ldots,y_N(t)|y_1(0),\ldots,y_N(0)) = \Pi_{i=1}^{N} P(y_i(t)|y_i(0))$

The Markov assumption for nucleotide substitution models implies that previous substitutions do not affect the probability of the next substitution, as seen from the first point in the above list. During the course of evolution the process of substitutions remains the same and shifting the same event to a later time does not change the probability of the event (the second in the list). The third property in the list shows that process of substitutions is the same for all the sites in the alignment. The last property is the assumption that the mutations in the columns of the alignment are independent of those in other columns.

With these 4 properties aforementioned, a 4-by-4 transition matrix can be made with the equations for the nucleotide substitution process in the phyloge-

netic trees,

$$\mathbf{P}(t) = \begin{pmatrix} P(y(t)=A|y(0)=A) & \dots & P(y(t)=A|y(0)=T) \\ P(y(t)=G|y(0)=A) & \dots & P(y(t)=G|y(0)=T) \\ P(y(t)=C|y(0)=A) & \dots & P(y(t)=C|y(0)=T) \\ P(y(t)=T|y(0)=A) & \dots & P(y(t)=T|y(0)=T) \end{pmatrix}. \tag{1.3}$$

The particular site where a mutation/substitution is made can be ignored since the process is identical along the sites as shown previously. The equations of the process in eq 1.3 use the symbol t as an indicator of time, please note that in other sections t is used for indexing the site number in the alignment. As mentioned, there is no chance for mutations to occur without a certain amount of time having passed and the identity matrix, $\mathbf{I}$, for a nucleotide substitution matrix arises when $t = 0$,

$$\mathbf{P}(t=0) = \mathbf{I} \tag{1.4}$$

The rate matrix, $\mathbf{Q}$, is a constant matrix. The values in the entries do not change over time, and is used to define the transition matrix for values of t other than 0. In subsection 1.3.2 the Jukes Cantor model is introduced and is where the concepts discussed here are demonstrated for the simplest cases. Later, the Kimura model (subsection 1.3.3) is shown with its increased flexibility and consequent increase in complexity.

For an infinitesimally small time interval dt the ansatz is made:

$$\mathbf{P}(dt) = \mathbf{P}(0) + \mathbf{Q}dt = \mathbf{I} + \mathbf{Q}dt \tag{1.5}$$

The homogeneous Markov chain satisfies the Chapman-Kolmogorov equation (Papoulis (1991) section 6 and pages 635-642):

$$\mathbf{P}(t_1 + t_2) = \mathbf{P}(t_1)\mathbf{P}(t_2) = \mathbf{P}(t_2)\mathbf{P}(t_1), \tag{1.6}$$

for arbitrary $t_1, t_2 \geq 0$. Setting $t_1 = t$ and $t_2 = dt$ the general expression for the transition matrix at any time t plus time dt is given by,

$$\mathbf{P}(t + dt) = \mathbf{P}(dt)\mathbf{P}(t). \tag{1.7}$$

Substituting $\mathbf{P}(dt)$ with that of eq 1.5 gives,

$$\mathbf{P}(t + dt) = (\mathbf{I} + \mathbf{Q}dt)\mathbf{P}(t) \tag{1.8}$$

which follows to produce;

$$P(t+dt) - P(t) = \mathbf{Q}\mathbf{P}(t)dt \tag{1.9}$$
$$\frac{P(t+dt) - P(t)}{dt} = \mathbf{Q}\mathbf{P}(t)$$

where $dt \to 0$ the left hand side can be rewritten,

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{Q}\mathbf{P}(t). \tag{1.10}$$

This differential equation has the solution,

$$\mathbf{P}(t) = e^{\mathbf{Q}t}, \tag{1.11}$$

and this gives the transition matrix for values of time t and is evaluated via a Taylor series expansion. Transition matrices must have their columns to sum to 1 to be a proper transition matrix. This is because there must be a valid distribution for the probability of observing a nucleotide at any time from all possible initial states. For this to arise the columns of the rate matrix must sum to 0, and is proved by (using eq 1.5):

$$1 = \Sigma_i P_{ik}(dt) = 1 + dt\Sigma_i Q_{ik} \Leftrightarrow \Sigma_i Q_{ik} = 0 \tag{1.12}$$

Here t is used to denote physical time, but in later sections the nucleotide substitution processes (and consequently the transition matrix) will be expressed in terms of *phylogenetic time* or the branch length, w. By defining $\lambda = 4\beta$ (where β is the rate of change for the substitutions),

$$w = \lambda t. \tag{1.13}$$

Very frequently we will express the probability of observing a nucleotide, y, as being dependent on the ancestral nucleotide, x, and the branch length w (the amount of phylogenetic time that has passed between the nucleotide present and the ancestral nucleotide); $P(y|x,w)$. The transition matrix is now rewritten with the conditional probabilities dependent on the ancestral nucleotide and the branch length as well,

$$\mathbf{P}(w) = \begin{pmatrix} P(A|A,w) & P(A|C,w) & P(A|G,w) & P(A|T,w) \\ P(G|A,w) & P(G|C,w) & P(G|G,w) & P(G|T,w) \\ P(C|A,w) & P(C|C,w) & P(C|G,w) & P(C|T,w) \\ P(T|A,w) & P(T|C,w) & P(T|G,w) & P(T|T,w) \end{pmatrix} \tag{1.14}$$

As pointed out in figure 1.4, for large branch lengths, the probability of each nucleotide is approximately $\frac{1}{4}$. This is the stationary distribution of the matrix which it converges to in the full extent of phylogenetic time and is invariant to the substitution matrix. It will be shown for this model that this has a uniform distribution over the nucleotides ($P(A) = P(C) = P(G) = P(T) = 0.25$) and is the equilibrium distribution for the transition matrix.

A column vector $\mathbf{u}$ dependent on the branch length (phylogenetic time) is used to represent the marginal distribution over the 4 possible nucleotides. This allows an investigation of the properties of the distributions of nucleotides and the branch lengths (over the course of time). For a general marginal distribution,

$$\mathbf{u}(w) = (P(y(w) = A), P(y(w) = C), P(y(w) = G), P(y(w) = T)) \tag{1.15}$$

and is a homogeneous Markov chain with the transition matrix $\mathbf{P}$ (as in eq 1.7),

$$\mathbf{u}(w_0 + w) = \mathbf{P}(w)\mathbf{u}(w_0). \tag{1.16}$$

and since the Markov chain is ergodic and converges to its stationary distribution regardless of the initial conditions. We can say that for the arbitrary set of nucleotides in $\mathbf{u}$,

$$\lim_{w\to\infty}\mathbf{u}(w) = \boldsymbol{\pi}, \tag{1.17}$$

where $\boldsymbol{\pi}$ is the stationary distribution (a vector of nucleotide probabilities). The vector $\boldsymbol{\pi}$ is denoted by,

$$\boldsymbol{\pi} = (\Pi_A, \Pi_C, \Pi_G, \Pi_T). \tag{1.18}$$

The invariance of the vector $\boldsymbol{\pi}$ towards the transition matrix gives the equilibrium distribution by,

$$\mathbf{P}(w)\boldsymbol{\pi} = \boldsymbol{\pi}. \tag{1.19}$$

The branch length is also a function of physical time, and is defined for all values of time greater than 0. The overall change is zero for the stationary equilibrium distribution, and therefore the rate matrix with respect to the equilibrium distribution is zero;

$$\mathbf{Q}\boldsymbol{\pi} = 0. \tag{1.20}$$

This is obtained by substituting eq 1.11 into eq 1.19.

From the property of *homogeneity*, the rate matrix is also assumed to be constant over the whole phylogenetic tree. For a given topology, each branch length of the branch length vector $\mathbf{w}$, has the same rate matrix and the equilibrium distribution of the nucleotides is the same.

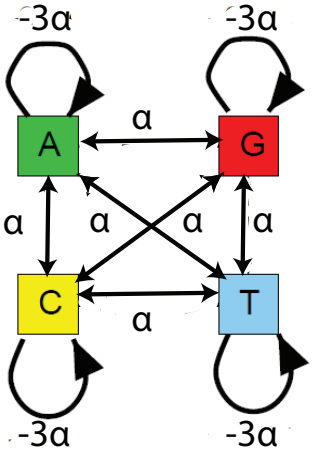


Figure 1.5: Jukes and Cantor model of nucleotide substitution

The nucleotide substitution where any different nucleotide replaces the original nucleotide occurs with a rate α. The non-diagonal elements in the rate matrix of eq 1.21 represent these substitutions with the rate α. The value -3α corresponds to the diagonal elements in eq 1.21 when the nucleotide doesn't change. The thickness in the arrows is proportional to the values of the substitutions' chances of occurring. Figure adapted from Husmeier *et al.* (2005a).

1.3.2 Jukes-Cantor model of nucleotide substitution

The Jukes-Cantor model of nucleotide substitutions is a special case of other models due to its simplicity, Jukes and Cantor (1969). The simplification is that it does not differentiate between transition and transversion substitutions, and has a uniform equilibrium distribution across the nucleotides. Every substitution from one nucleotide to another nucleotide of different value is treated equally. The rate matrix $\mathbf{Q}$ for this model is,

$$\mathbf{Q} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}. \tag{1.21}$$

From the matrix the rate of substitution to different nucleotides is α and substitutions not changing the nucleotide is -3α. Figure 1.5 displays a diagram of the Jukes Cantor substitution process.

For the 2 different types of substitution events the transition matrix will use the notation for substitution events into the same nucleotide with $\tilde{d}(t)$ and the

transitions into different nucleotides with $\tilde{g}(t)$),

$$\mathbf{P}(t) = \begin{pmatrix} \tilde{d}(t) & \tilde{g}(t) & \tilde{g}(t) & \tilde{g}(t) \\ \tilde{g}(t) & \tilde{d}(t) & \tilde{g}(t) & \tilde{g}(t) \\ \tilde{g}(t) & \tilde{g}(t) & \tilde{d}(t) & \tilde{g}(t) \\ \tilde{g}(t) & \tilde{g}(t) & \tilde{g}(t) & \tilde{d}(t) \end{pmatrix}. \tag{1.22}$$

Eq 1.10 defines the rate of change for the transition matrix, ${}^{d\mathbf{P}(t)}/_{dt} = \mathbf{RP}(t)$; for the Jukes-Cantor model the elements of the rate matrix can be substituted giving:

$$\frac{dP_{ij}(t)}{dt} = -3\alpha P_{ij}(t) + \alpha \sum_{k \neq j} P_{ik}(t). \tag{1.23}$$

The term for the sum of the non-identical substitutions can be simplified by using $\sum_{k \neq j} \mathbf{P}_{ik}(t) = 1 - P_{ij}$ and inserted into the above,

$$\frac{dP_{ij}(t)}{dt} = -3\alpha P_{ij}(t) + \alpha \left(1 - P_{ij}\right) = \alpha - 4\alpha P_{ij}(t). \tag{1.24}$$

This linear differential equation when solved requires the initial conditions $\mathbf{P}_{ii}(0) = 1$ meaning that the probability of there being no mutation at time 0 is 1 and for non-identical substitutions where $i \neq j$ the probability is 0, $\mathbf{P}_{ij}(0) = 0$. A solution is found for the identical and non-identical substitutions as follows:

$$P_{ii}(t) = 0.25 + 0.75e^{-4\alpha t} \tag{1.25}$$

$$P_{ij}(t) = 0.25 - 0.25e^{-4\alpha t}. \tag{1.26}$$

As time approaches infinity the second term in both equations becomes negligible and the probability for both equations and therefore each nucleotide is the same at 0.25 which is the stationary distribution. This shows how the matrix entries in eq 1.22 and eq 1.27 can be found. The equations for the entries of substitution events are (utilising eq 1.11 as when producing eq 1.45- 1.47 and the previous equations),

$$\tilde{f}(t) = \frac{1}{4}\left(1 - e^{-4\alpha t}\right) \tag{1.27}$$

$$\tilde{d}(t) = 1 - 3\tilde{f}(t) = \frac{1}{4}\left(1 + 3e^{-4\alpha t}\right). \tag{1.28}$$

For this model the free parameter is α which is not identifiable (confounded with t). The value of α is chosen such that the branch lengths indicate the average number of mutations. The general case is found in Minin *et al.* (2005),

$$\sum_{j} \mathbf{Q_{ij}} \pi = -1. \tag{1.29}$$

The symmetry of the Jukes-Cantor rate matrix allows a scalar form to be derived; $4(-3\alpha)^1/_4 = -1 \Rightarrow 3\alpha = 1$. This involves using a variable to denote this, $\lambda = 3\alpha$. These events are modelled by a Poisson process (in time t) with decay rate $\lambda = 3\alpha$

,

$$P_j(t) = \frac{e^{\lambda t}(\lambda t)^j}{j!}, \quad j = 1,2,3 \tag{1.30}$$

and here j represents the number mutations. Setting the branch lengths to be $w = \lambda t$, the equation is,

$$P_j(w) = \frac{e^{-w}w^j}{j!}. \tag{1.31}$$

For the average number of mutations $< j >$ the following can be derived,

$$\begin{aligned} < j > \quad &= \sum_0^\infty j P_j(w) && (1.32)\\ &= \sum_{j=0}^\infty j \frac{e^{-w}w^j}{j!} && (1.33)\\ &= e^{-w}\sum_{j=1}^\infty \frac{w^j}{(j-1)!} && (1.34)\\ &= e^{-w}w\sum_{j=1}^\infty \frac{w^{j-1}}{(j-1)!} && (1.35)\\ &= e^{-w}w\sum_{j=0}^\infty \frac{w^j}{(j)!} && (1.36)\\ &= e^{-w}we^w && (1.37)\\ &= w. && (1.38) \end{aligned}$$

To interpret the evolutionary process in terms of branch lengths, $w = \lambda t$, we can define $\lambda = 3\alpha$ and the equations become:

$$f(w) = \frac{1}{4}\left(1 - e^{-\frac{4}{3}w}\right) \tag{1.39}$$

$$d(w) = 1 - 3\tilde{f}(w) = \frac{1}{4}\left(1 + 3e^{-\frac{4}{3}w}\right). \tag{1.40}$$

This allows us to easily substitute these equations in eq 1.22 to get $\mathbf{P}(w)$. The model has the equilibrium distribution as $t \to \infty$ and the stationary vector π:

$$\mathbf{Q}\pi = 0, \tag{1.41}$$

is

$$\pi = \left(\Pi_A = \frac{1}{4}, \Pi_C = \frac{1}{4}, \Pi_G = \frac{1}{4}, \Pi_T = \frac{1}{4}\right). \tag{1.42}$$

This equilibrium distribution can be found easily from examining the limiting cases of eq 1.22 and eq 1.39. There is only 1 free parameter in this model being the value of α, which under the setting for the branch lengths results in no free parameters. α is not identifiable and which is set such that the branch lengths

can be interpreted as the average number of mutations. The Kimura model of the next section 1.3.3 becomes equivalent to this model when the α and β parameters are equal to each other.

1.3.3 Kimura Model

The Kimura model Kimura (1981) is a more flexible model of nucleotide substitution than the Jukes Cantor model previously described in subsection 1.3.2. The Kimura rate matrix has the form,

$$\mathbf{Q} = \begin{pmatrix} -2\beta-\alpha & \beta & \alpha & \beta \\ \beta & -2\beta-\alpha & \beta & \alpha \\ \alpha & \beta & -2\beta-\alpha & \beta \\ \beta & \alpha & \beta & -2\beta-\alpha \end{pmatrix} \tag{1.43}$$

The rows from top to bottom and the columns from the left to right correspond to the nucleotides in the order that they do in eq 1.3 (A/C/G/T). The α parameter in the rate matrix denotes the rate for a transition, and β that of a transversion. Subsection 1.5 introduced the difference between these two types of substitutions. The figure 1.6 depicts the model via a diagram showing the rates of substitutions between the 4 possible nucleotides. The rate of substitution is not uniform across all possibilities as it is for the Jukes-Cantor model shown in figure 1.5. The thickness of an arrow is approximately proportional to the rate of substitution between nucleotides. This is used to depict how the transition substitutions are more frequent than the transversion substitutions.

The 3 different types of substitution events in the transition matrix are abbreviated; the substitution into the same nucleotide with $\tilde{d}(t)$, the transitions with $\tilde{g}(t)$, and the transversions with $\tilde{f}(t)$:

$$\mathbf{P}(t) = \begin{pmatrix} \tilde{d}(t) & \tilde{f}(t) & \tilde{g}(t) & \tilde{f}(t) \\ \tilde{f}(t) & \tilde{d}(t) & \tilde{f}(t) & \tilde{g}(t) \\ \tilde{g}(t) & \tilde{f}(t) & \tilde{d}(t) & \tilde{f}(t) \\ \tilde{f}(t) & \tilde{g}(t) & \tilde{f}(t) & \tilde{d}(t) \end{pmatrix}. \tag{1.44}$$

The equations for these entries of substitution events are (which were derived for

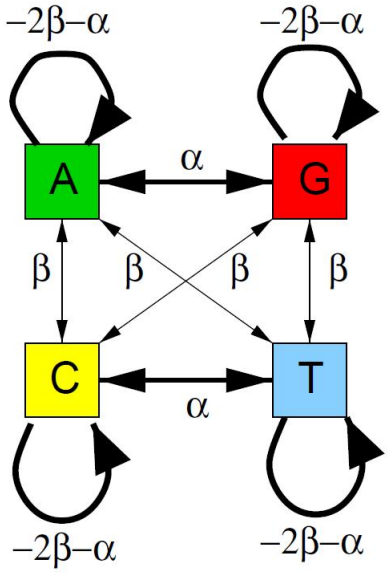


Figure 1.6: Kimura model of nucleotide substitution

The parameter for the transitions is denoted with α, and for the transversions with β. The value $-2\beta-\alpha$ corresponds to the diagonal elements in eq 1.43 denoting events where the nucleotide does not change. The thickness in the arrows is proportional to the values of the substitutions' chances of occurring and it is seen how the transitions are more likely than the transversions. Figure taken from Husmeier *et al.* (2005a).

the Jukes-Cantor model),

$$\tilde{f}(t) = \frac{1}{4}\left(1 - e^{-4\beta t}\right) \tag{1.45}$$

$$\tilde{g}(t) = \frac{1}{4}\left(1 + e^{-4\beta t} - 2e^{-2(\alpha+\beta)t}\right) \tag{1.46}$$

$$\tilde{d}(t) = \left(1 - 2\tilde{f}(t) - \tilde{g}(t)\right), \tag{1.47}$$

which arise from eq 1.11 and can be found in Durbin *et al.* (1998). Now the *transition-transversion ratio* can be defined in the proper context,

$$\tau = \frac{\alpha}{\beta}. \tag{1.48}$$

Substituting the branch lengths into eq 1.45- 1.47 gives,

$$f(w) = \frac{1}{4}\left(1 - e^{-\frac{4}{3}(2+\tau)w}\right), \tag{1.49}$$

$$g(w) = \frac{1}{4}\left(1 + e^{-\frac{4}{3}w} - 2e^{-\frac{4}{3}\frac{\tau+1}{2}w}\right), \tag{1.50}$$

$$d(w) = 1 - 2f(\frac{4}{3}w) - g(\frac{4}{3}w). \tag{1.51}$$

This model has 2 free parameters since the stationary distribution is fixed at the uniform distribution. From the transition-transversion ratio, τ, there are 2 free parameters, the α and β, minus 1 (for identifiability) leaving 1 free parameter for this model. The transition-transversion parameter is usually set to the default value, $\tau = 2$, producing the graphs of the demonstration of nucleotide models in fig 1.4. The actual number of free parameters will be minus 1 of this number due to the constraint that parameter β is fixed to solve an identifiability issue for the transition transversion ratio, as written in Minin *et al.* (2005). The identifiability issue arises between the magnitudes of α and β taking on multiple values for the same value of τ. The constraint used is $\sum_i \mathbf{R}_{ii}\pi_i = -1$ which was solved for the scalar situation of the Jukes-Cantor model in eq 1.23 and eq 1.24.

A more complex method, the HKY model, is described in the appendix in section A.3. It allows greater flexibility in modelling substitutions by having a non-uniform stationary distribution for the base frequencies of the nucleotides.

1.4 Non-probabilistic methods of phylogenetic tree reconstruction

This thesis is concerned with probabilistic methods of phylogenetic tree reconstruction, with the main motivation being that the process of evolution is in-

herently stochastic producing mutations at random. There are 3 main groups of methods of phylogenetic tree construction in general use, which are based on genetic distance, clustering and parsimony. A short introduction to these non-probabilistic methods is presented later in this section.

The main benefit of non-probabilistic methods is that they can be implemented with ease and they have very efficient completion times. For large groups of sequences, such as the trees of life, non-probabilistic methods are commonly used for grouping together hundreds of species in large sequence alignments. Dirk Husmeier (2003) reviews many of these methods in chapter 4 of his book.

1.4.1 Evolutionary distances and clustering

All distance based or clustering methods of sequence alignments have in common that they do not consider the sequences in relation to a phylogenetic tree. Instead these methods build the tree iteratively based upon the distance metric as each of the sequences are processed. Having a distance metric does not assist in constructing ancestral species. A benefit of this method over that of parsimony (subsection 1.4.2) and maximum likelihood (section 1.5) is that the order of the sequences in the alignment does not affect the results. The main drawback of these methods is that they do not account for substitutions which are not observable from the extant sequence data, and that they suffer from information loss. These two consequences are visible from the two images shown in figure 1.7. Subfigure a) shows various situations that may arise where the true number of substitutions is not estimated using a distance measure between sequences. Subfigure b) shows how a distance matrix can be constructed from the pairwise distances between sequences. Such a matrix contains less information than a phylogenetic tree, as it is additionally capable of representing the order of the ancestry and the type of the substitutions.

1.4.2 Parsimony

Parsimony searches for the evolutionary history which minimises the number of mutations between sequences. The columns (sites) of the sequence alignment are considered to be independent and a candidate topology for the alignment is chosen when assessing its suitability to the data. For each topology considered, the number of mutations required between the ancestral species and the extant

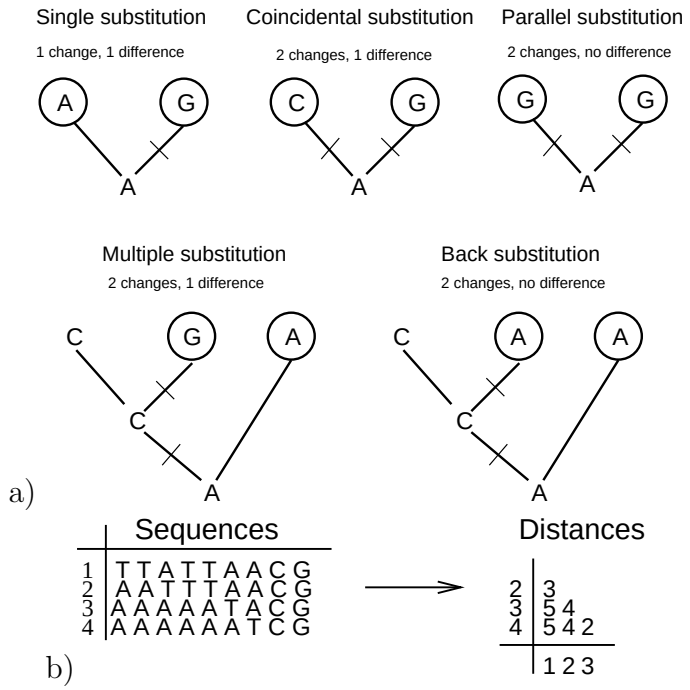


Figure 1.7: Depiction of the 2 main drawbacks of distance methods

Subfigure a) shows how using pairwise distances of the sequence data can underestimate the true number of substitutions. Subfigure b) shows how distance metrics on sequence data results in information loss compared to methods which construct a phylogenetic tree. Figures are taken from Husmeier *et al.* (2005a).

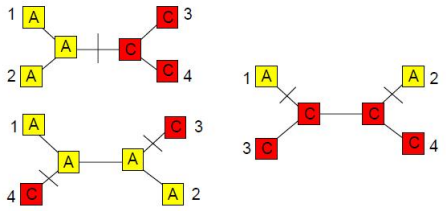


Figure 1.8: Parsimonious phylogenetic tree reconstruction

The figure shows how a topology is chosen using the parsimonious method of phylogenetic tree construction. This approach is demonstrated using trees constructed from sequence alignments of 4 taxa. In this example the column of nucleotides is, A,A,C,C. From the 3 candidate topologies we can see that the first one grouping taxa 1 and 2 to be adjacent to each other has only 1 required mutation along the central branch length whereas the other two topologies require 2 mutations each. Parsimony therefore would chooses the first topology. Figure taken from Husmeier *et al.* (2005a).

species (more accurately the minimum the number of mutations required for the transition) is calculated. The topology which necessitates the fewest substitutions over all the sequence alignment is the chosen topology.

Figure 1.8 demonstrates how phylogenetic trees are reconstructed using parsimony. The example data that the method is given is a sequence alignment of 4 taxa that have the nucleotides A, A, C, C at a particular site. From the 3 possible topologies that can be created using 4 taxa, the first topology grouping taxa 1 and 2 to be adjacent has only 1 mutation whereas the other 2 topologies have 2 mutations and are therefore less optimal given the criteria of the method to minimise the number of mutations. The ancestral base pairs are chosen so as to minimise the number of mutations for the topology. Each topology has an associated count of the total number of mutations it incurs along the whole alignment. The topology with the lowest count of mutations is chosen.

1.4.3 Felsenstein zone

Long branch attraction is an effect which causes methods such as parsimony 1.4.2 to infer the incorrect topology regardless of the amount of data provided. The failure is inherent to the model, and the situations in which phylogenetic trees produce data susceptible to this failure, are termed as being in the *Felsenstein*

zone, Felsenstein (1978a).

Long branch attraction is seen when two taxa from different bifurcations on the phylogenetic tree (non-adjacent taxa) have branch lengths proportionally larger than the taxa they are grouped with (adjacent taxa). The lengths of non-adjacent long branches are not taken into account in the nucleotide substitution model and so it fails to relax the penalisations from having different nucleotides with the adjacent taxa (which occurs frequently). As a consequence, a higher scoring tree (one counting fewer substitutions) is found which groups the non-adjacent taxa. This is best described by the subfigure a) in figure 1.9 which presents the incorrect inference performed by parsimony. In the example, the strains with long branches have different nucleotides compared to their closest relatives, and parsimony groups them together to minimise the penalties along the topology. In the probabilistic framework, on the other hand, these mismatches incur smaller penalisations allowing them to have a smaller influence on the chosen topology.

Felsenstein (1978a) derives (for the general case) situations where models based on parsimony tree construction methods will consistently infer the incorrect topology. This is shown for all cases with 4 sequences (resulting in trees with 5 branch lengths) where 2 non-adjacent branch lengths share the same length d_2, and the rest of the 3 branch lengths share an equal length d_3 independent of d_2. A region where correct and incorrect inference occurs is shown for these two lengths in subfigure b) of figure1.9. For both sets of lengths, d_2 is on the x-axis and d_3 on the y-axis. Label 'C' stands for 'correct' inference and 'NC' for 'incorrect' inference. In region 'C' for the those values of d_2 and d_3, the topology grouping the long branches is chosen rather than the topology used to generate the data. Chapter 2 investigates these regions of correct and incorrect inference using data sets spanning these 2 regions.

1.5 Likelihood methods

This section states some of the basic motivations and mathematical foundations for probabilistic approaches towards building phylogenetic trees. The difference between a statistical model and the statistical inference is that the model describes the object of concern by setting the equations of the variables and probability distributions to the observations seen. Statistical inference studies how the random samples observed are used in finding the parameters of the model.

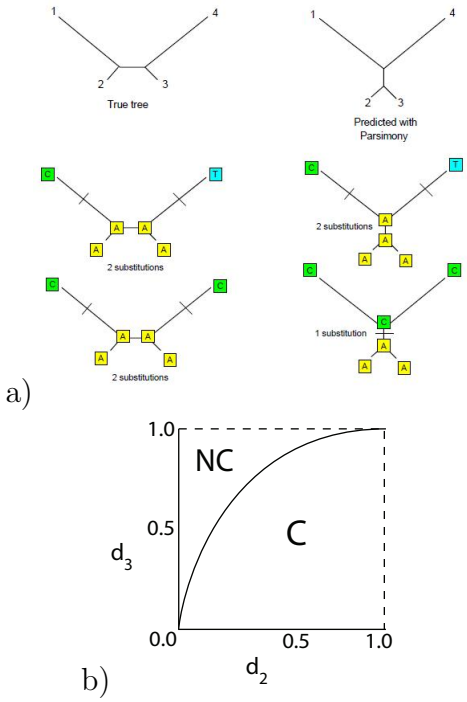


Figure 1.9: The consistent/non-consistent regions for tree estimation with parsimony Subfigure a) shows how an application of parsimony can result in inferring the incorrect topology if the tree falls in the Felsenstein zone (Felsenstein (1978b)). In situations where two non-adjacent taxa both have substantially larger branch lengths than their adjacent taxas' branch length, they are susceptible to long branch attraction. In subfigure b), for the two groups of branch lengths d_2 and d_3 the regions of (c)onsistent and (n)on-(c)onsistent estimation of the true underlying phylogenetic tree is shown. For a tree of 4 strains d_2 includes the middle branch lengths and two non-adjacent branches, an d_3 represents the 2 remaining non-adjacent branches. 'C' denotes consistency where when the branch lengths of the two groups have a ratio within this region, then the parsimonious estimation of the underlying tree is consistent with the data generating process. 'NC' denotes the region where the estimation process is not consistent with the underlying true phylogeny. Subfigure a) taken from Husmeier *et al.* (2005a), and subfigure b) adapted from Felsenstein (1978b).

Likelihood methods are founded on a mathematical model that is defined explicitly. It does not take an approach where it implies a certain model of operation. This is beneficial since the assumptions made can be examined, put under scrutiny and be modified more easily. Likelihood, which probabilistic methods use to optimise a clear criteria, are similar in this respect to parsimony but different to clustering methods. Parameters for features of evolution inspired from biological evidence can also be incorporated into likelihood models.

A phylogenetic tree's likelihood can be computed by considering the tree structure as a Bayesian network which is a graphical model. This thesis considers only unrooted trees, and in computing the likelihood of the unrooted tree, the parent node of the network can be placed anywhere along the ancestral nodes of the graph. The directions of edges in the tree must be determined by an arbitrarily placed root node for the Bayesian network to be factorised into a product of conditional probabilities between the ancestral and extant nodes. Having a substitution model based on likelihood as shown in figure 1.4 allows a probability associated with substitutions between nodes in the network to be computed, and a joint likelihood from the factorisation can be made.

Figure 1.10 shows in subfigure a) the directed graph of the Bayesian network. Any general phylogenetic tree structure with a certain number of nodes can be always be factorised to allow computation of the likelihood. Nucleotides are represented by variable x, can take the value of any nucleotide $x_i \in A,C,G,T$. Arrows on the edges indicate descendants of nodes that are ancestors of the node from where the arrow emerges from. These parent/ancestral nodes can be identified with a subscript $pa[i]$. This way every nucleotide is descended from another nucleotide which then carries back the probability of the root node to be the ancestral nucleotide of all the nodes in the network. For a tree arising from m sequences belonging to m different organisms, the factorisation of the likelihood is given by:

$$P(x_1,\ldots,x_m) = \prod(x_{root}) \prod_{i \in T/r} P(x_i|x_{pa[i]},w_i). \tag{1.52}$$

In expressing the set of the nodes in the network, T is introduced here to denote the set of all the nodes in the tree. The product of all the conditional probabilities in the tree excluding that of the root which is separate is denoted by $i \in T/r$. The probability of the nucleotides at the root can be set to any normalised distribution and is taken to be uniform. In order to calculate the probabilities of the observed

nucleotide sequence we must make an assumption about the root having a certain nucleotide distribution. We are also compelled to assume that the ancestral sites have this same distribution for the nucleotides at any site in the alignment. From figure 1.10, subfigure a) shows the ancestral nodes shaded in lightly (z variables as unknowns) and those parts with darker shading are the observed nucleotide positions in the sequence alignment (extant nucleotides with y). The subscript of the node refers to which of the known or unknown nodes it refers to. The joint probability for the nodes of the network, $P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S)$ is,

$$\begin{aligned} P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S) = \\ P(y_1|z_1, w_1)P(y_2|z_1, w_2)P(z_2|z_1, w_5)P(y_3|z_2, w_3)P(y_4|z_2, w_4)\Pi(z_1), \end{aligned} \tag{1.53}$$

where z_1 is chosen to be the root node. The root node, $\Pi(z_1)$, has probability $\frac{1}{4}$ for all the nucleotides. The other components are the conditional probabilities defined by the substitution matrix which is introduced in figure 1.4.

Subfigure b) in figure 1.10 shows the 16 possible nucleotide values the ancestral nodes can take. They are marginalised over since they are not observed. The summations of the hidden nodes in eq 1.53 can be done straight forwardly with nested summations over the two hidden node variables, z_1.

$$P(\mathbf{y}|\mathbf{w}, S) = P(y_1, y_2, y_3, y_4|\mathbf{w}, S) = \Sigma_{z_2}\Sigma_{z_1}P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S), \tag{1.54}$$

shows the marginalisation over the hidden ancestral nodes and obtains the probability of the extant nucleotides. For an unrooted tree of m taxa (leaves) there are $m-2$ ancestral unobserved nodes. In general there are 4^{m-2} computations to be performed in the marginalisation over all these nodes. To avoid this computational demand required for larger alignments, the *peeling algorithm* of Felsenstein (1981) is used. It uses the sparse connectivity of the network, that there are only bifurcations at the edges and the independence of nodes conditional on each other. The exponential running time can be made polynomial similarly to Pearl's *message passing algorithm*, Pearl (1988).

The difference between rooted and unrooted trees is that for unrooted trees that the root node can be chosen to be anywhere on the tree and for rooted trees it has a fixed position. This is not discussed further in this thesis as only unrooted trees are used. Unrooted trees use reversible models of nucleotide substitution as defined in the Kimura model described in section 1.3.3. When calculating the likelihoods with a reversible model of nucleotide substitution, the probabilities

between the substitutions and the reverse substitution, conditional that the same amount of phylogenetic time has passed (the same branch length) produces equal probabilities:

$$P(x|y,w)\Pi(y) = P(y|x,w)\Pi(x). \tag{1.55}$$

The variables x and y take on the value of one of the nucleotides. Therefore the direction of the edges on the Bayesian network does not change the probability. Changing the position of the root (that changes the edge directions along some of the branch lengths) will not change the probability of the network. The root distribution of nucleotides applies to the whole likelihood so that the overall probability is the same irrespective of the root position. This is explained in section 4.3 of Husmeier *et al.* (2005a).

At each site in an alignment of N columns, there are m nucleotides (gaps are rejected), and the columns are addressed as $\mathbf{y}_t$. t is used as an index along the alignment $1 \leq t \leq N$. The peeling algorithm together with a nucleotide substitution model allows the probability of a column of nucleotides in the sequence alignment to be computed,

$$P(y_t|\mathbf{w},S). \tag{1.56}$$

To compute the likelihood for the whole sequence alignment it is assumed that the columns are independent of each other. The N sites of the alignment make the data used, $\mathcal{D} = \mathbf{y}_1,\ldots,\mathbf{y}_N$, and the independence assumption allows the factorisation of the whole alignment,

$$P(\mathcal{D}|\mathbf{w},S) = \prod_{t=1}^{N} P(\mathbf{y}_t|\mathbf{w},S). \tag{1.57}$$

Increasing the likelihood of the data is the objective of the inference schemes applied. This model will be changed in later sections, as will be shown. This is the basic model which further modifications are built upon. DNAML is a maximum likelihood implementation which is openly available, Felsenstein (1981) and Felsenstein (1996).

1.6 Recombination

More accurately, this subsection discusses *inter-specific recombination*, which will be referred to as *recombination*. Recombination in the study of phylogenetics has been important for the correct reconstruction of the evolutionary histories of

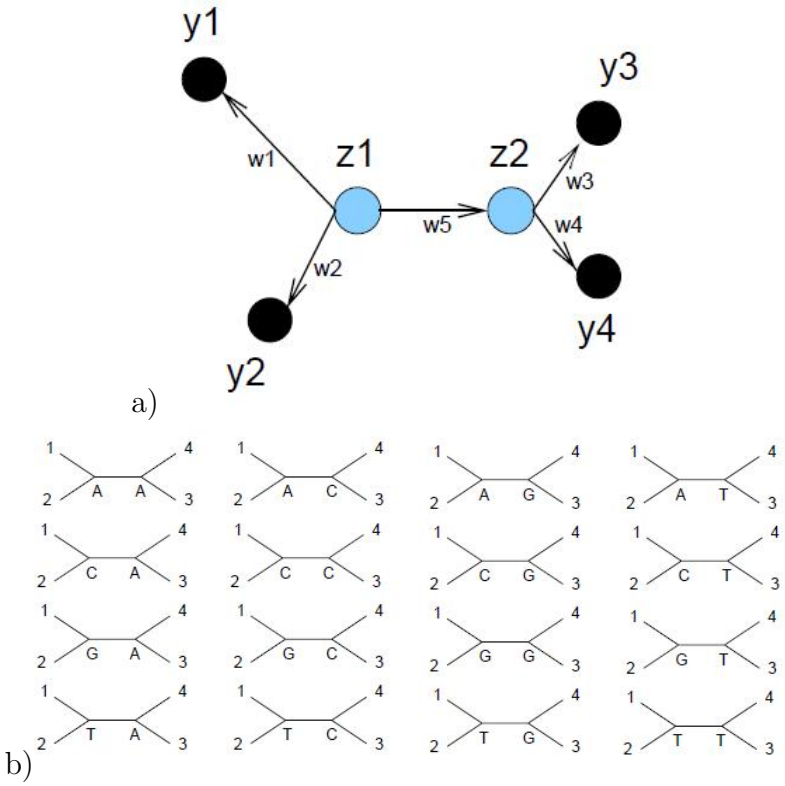


Figure 1.10: Graphical Model of the Phylogenetic tree

Subfigure a) shows a Bayesian network of a phylogenetic tree. The lightly shaded nodes represent the ancestral species of the extant species which are observed, and the extant observed data are black nodes. The arrows show the direction of the dependencies between the nodes, the children of nodes are indicated by the arrowheads. Subfigure b) shows the 16 possible different settings for the 2 ancestral unobserved nodes in the 4 taxa phylogenetic tree. These are marginalised over to obtain a likelihood for the topology and other evolutionary parameters such as the ratefactor and branch length vector. (figures taken from talks of Dirk Husmeier)

many unicellular pathogens. This relatively recent interest is less developed and is not considered to be a part of the traditional approach towards phylogenetics.

The introduction into phylogenetics till this point discussed finding a single phylogenetic tree (topology) describing the whole sequence alignment's evolutionary history. The process of recombination results in the change of the topology along an alignment. This provides a more detailed explanation of the ancestry of the present genetic material. With recombination events there can be two or more topologies along the alignment creating a *mosaic structure*.

It is important to detect and infer recombination events and where they occur to produce correct phylogenetic relationships for the data. Wrongly attributed trees may result in false conclusions from investigations. Accounting for recombination adds a significant amount of complexity to the model in terms of parameters that need to be assumed or estimated in the model. The number of topologies scales super exponentially with the number of sequences in the alignment and is a large obstacle in phylogenetics. Modelling recombination involves positioning break points between choices of topology along the alignment, each of these topologies must also be inferred, and so modelling recombination creates a difficult problem. To further exasperate the problem, the placement of the break points is not an analytically tractable problem and is combinatorical in its nature.

1.6.1 The recombination process

Recombination events happen in unicellular organisms such as bacteria and viruses. They can exchange or transfer genetic material (DNA subsequences) between themselves. These subsequences may also be referred to as *mosaic* sequences.

Figure 1.11 depicts the process of recombination and the effect that it has on the phylogenetic trees that are produced. The process follows the steps that two non-adjacent strains on the phylogenetic tree exchange a subsequence of their DNA between themselves. In the regions where there has not been an exchange the phylogenetic relationships between the strains remains unchanged. In the regions where an exchange has happened, the phylogenetic relationship (topology) changes to bring together strains that contain a more similar evolutionary history (closer in terms of phylogenetic time shown in the branch lengths). The figure shows a region of horizontal transfer, the consequences it has on the sequence

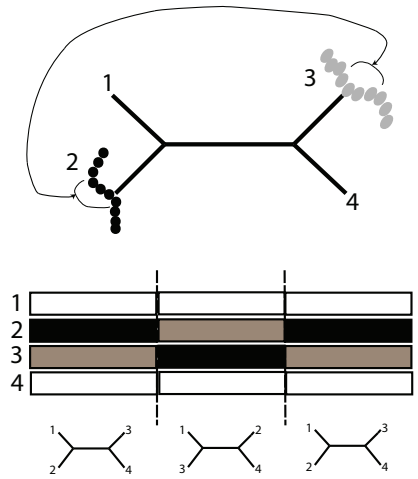


Figure 1.11: Recombination in DNA sequence alignments

The figure shows a phylogenetic tree where strains 2 and 3 undergo a recombination event in the central third of their genome. The transfer of genetic material is seen in the change of darkness in the shading of strains 2 and 3 in the central region. The regions of DNA surrounding the recombination event have the same phylogenetic tree as the one on the top as no change has occurred there. In the region where the recombination event occurred, a change in the topology is required so that the correct branching order is found that brings sequences together with their closest relationships in adjacency. In the central region, the sequence labelled 2 contains genetic material of strain 3 originally which is closest to strain 4, and the topology in this region changes to group those 2 strains together.

alignment, and the effect on the correctly inferred topologies in the respective regions.

1.6.2 Classical methods for detecting recombination

There are various methods which can be used for detecting recombination that have different approaches and features. *Phylogenetic networks* is one method which can be used as an indicator for recombination in sequence alignments as well as indicating other types of events. The method builds cyclic networks based on probabilistic or non-probabilistic measures. The set of possible supported trees are represented by *splits* without constructing ancestral species. The prediction of recombination events is made without an estimate of the specific sites where this occurs. This method is described in more detail in section 5.1 in

chapter 5 where it is applied. Maximum chi-square, Maynard Smith (1992), is another method which compares pairs of sequences with a putative recombinant to detect recombination. This method does not take into account explicit phylogenetic relationships, and only takes into account polymorphic sites in DNA which does not utilise the maximum amount of possible information. This results in a poor resolution of the recombination break points. *Partial Likelihoods Assessed Through Optimisation (PLATO)*, which is proposed in Grassly and Holmes (1997) compares average likelihoods of putative recombinant regions against the average likelihood of the whole sequence. It uses a statistic to determine topology changes and the method becomes increasingly unreliable as the length of the recombination region grows. TOPAL is another method which is used in the chapter 5 and described in section 5.2. It is a fast method which can be applied to large sequence alignments but suffers from the information loss inherent to distance methods. There is also a method of detecting recombination based on parsimony, RECPARS, and is presented in Hein (1993). It requires from the user a set of tuning parameters to be chosen and suffers from long branch attraction in the Felsenstein zone.

1.7 Exploring the effect of rate heterogeneity and recombination across sequence alignments

The rate of mutation, which scales the branch lengths in a phylogenetic tree is not uniform across DNA sequence alignments. Heterogeneity in the rate of mutation can occur for a variety of reasons and has significant biological importance. Inferring the correct rate of mutation for the different regions of an alignment has also implications for inferring topology break points reliably as will be discussed in later sections.

Figure 1.12 has two subfigures to demonstrate the changing of the rate of mutation (via a ratefactor) and the branch length vector. Subfigure a) shows three regions of differing rate heterogeneity. The sections show a ratefactor equal to 1, less and greater than 1. The topology is kept the same, as well as the branch length vector, but the branch lengths are all scaled according to the ratefactor uniformly along the alignment. Subfigure b) shows 3 different trees of the same topology with differing branch length vectors. The ratefactor value is kept con-

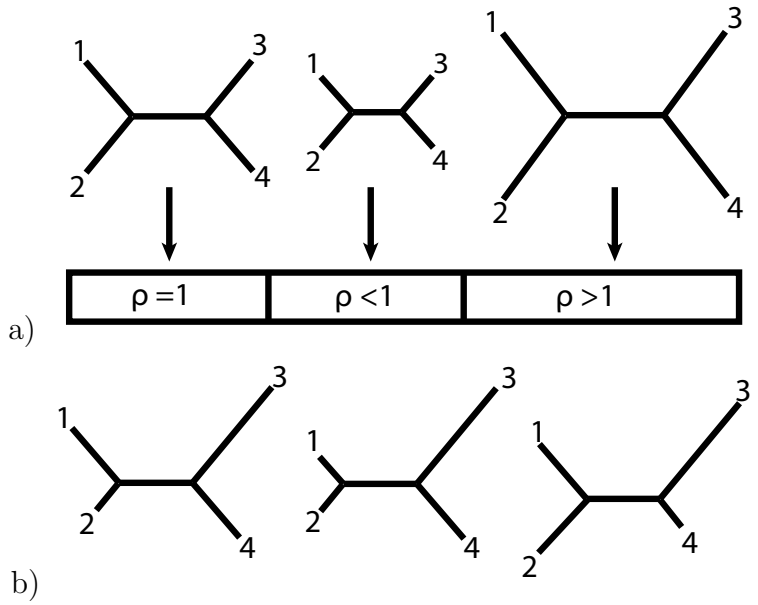


Figure 1.12: The effect of rate heterogeneity and changes in branch lengths to phylogenetic trees

The two subfigures depict the effects of rate heterogeneity and branch length differences. Subfigure a) shows three regions of rate heterogeneity on a sequence alignment. The regions have a ratefactor equal, less and greater than one. The topology stays the same and so does the branch length vector along the sequence. The number of expected mutations observed are scaled with the ratefactor. The branch lengths vary according to ρ. Subfigure b) demonstrates three different trees of the same topology and ratefactor, but having a different branch length vector.

stant meaning that the relative values change (direction of the vector of branch lengths), but the sum of the lengths $\sum \dot{w}_i = constant$ over the three different trees.

As was illustrated in Figure 1.11, there are topology breakpoints along the DNA sequence alignment as well. The break points for rate heterogeneity can coincide with those of the topology break points, but this is not a requirement (since real biological data can produce these effects independently of one or the other). Figure 1.13 illustrates this. Subfigure a) shows the more specific case where break points between the different ratefactors coincide with those for the topology changes along the alignment. Subfigure b) shows an example alignment where the ratefactor breakpoints do not coincide with those of the topology break points. It is seen that the taxa adjacencies in the phylogenetic trees attributed

to different regions can change and within homogeneous topology regions the scaling for the complete set of branch lengths can change as well and this rate heterogeneity can be maintained over further topology changes.

The independent detection of topology and ratefactor break points is what is achieved in Husmeier (2005) with the implementation of the phylogenetic factorial HMM (phylo-FHMM). Subsection 1.9.4 describes the model used to infer and place the break points for the two factors separately.

1.8 Bayes Theorem and Bayesian Networks

Throughout this thesis, we will use the Bayesian paradigm as it allows to quantify uncertainties and incorporate knowledge in a natural way. A prior probability is needed to express a degree of belief in events before measured data appears, a likelihood function that models the data, and this gives an estimate of the uncertainty. When considering a certain hypothesis H the concern is with the posterior probability of the hypothesis given the data, which is proportionate to the product of the prior probability and the likelihood.

Conditional probabilities for H the hypothesis and $\mathcal{D}$ the data are;

$$P(\mathcal{D}|H) = \frac{P(\mathcal{D},H)}{P(H)} \tag{1.58}$$

and

$$P(H|\mathcal{D}) = \frac{P(\mathcal{D},H)}{P(\mathcal{D})}. \tag{1.59}$$

Bayes formula is based on the *prior* probability of the event/hypothesis we are concerned with, the *likelihood* of the data under the event of the hypothesis, and the *marginal* likelihood of the data. These allow the *posterior* probability of the hypothesis given the data to be computed,

$$\text{posterior} = \frac{\text{prior x likelihood}}{\text{marginal likelihood}} \tag{1.60}$$

$P(H)$ is the prior probability for event H, $P(\mathcal{D}|H)$ is the likelihood, $P(\mathcal{D})$ is the marginal likelihood of the data, and $P(H|\mathcal{D})$ the posterior probability;

$$P(H|\mathcal{D}) = \frac{P(H)P(\mathcal{D}|H)}{P(\mathcal{D})}. \tag{1.61}$$

The hypothesis can concern many different choices such as the dependency structure between sets of proteins in a genetic network, the association between smoking and cancer, the topology state sequence along a DNA sequence alignment, and many others.

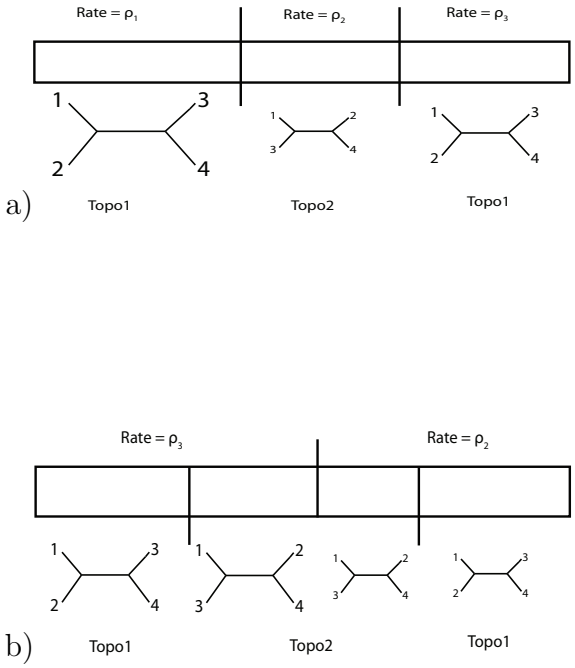


Figure 1.13: Illustration of the effect on a phylogenetic tree in the presence of rate heterogeneity as well as topology break points

This figure has two subfigures a) and b) which illustrate the effect on the phylogenetic tree along the DNA sequence alignment when there are both topology break points arising from recombination as well as break points for the rate heterogeneity. 'Topo' is used for the abbreviation of topology. Topology 1 joins strains 1 and 2 to be adjacent, and Topology 2 has strains 1 and 3 being adjacent. There are 3 rate state values used ρ_1, ρ_2, and ρ_3 where $\rho_2 < \rho_3 < \rho_1$. Subfigure a) shows two break points separating the alignment into 3 sections. Before the first and after the second breakpoint the regions have the same topology and the change is that the central region contains Topology 2. Each region has a different ratefactor uniformly scaling the complete set of branch lengths. Subfigure b) differs from a) in that there is now only 1 break point for the ratefactors in the central region. This does not align with a break point for the topologies. As a result, in the region where a recombination event has occurred, the scaling of the branch lengths changes in this region.

The marginal likelihood of the data $P(\mathcal{D})$ is the sum of the space of different hypothesis which are independent without overlap between them (it is the denominator in the partition function acting as a normalising constant),

$$P(\mathcal{D}) = \Sigma_i P(\mathcal{D}, H_i) = \Sigma_i P(\mathcal{D}|H_i)P(H_i), \tag{1.62}$$

where i is used here as an index for the individual hypothesis. Since $P(\mathcal{D})$ is a constant between all the hypothesis the comparison can be written as

$$P(H|\mathcal{D}) \propto P(H)P(\mathcal{D}|H), \tag{1.63}$$

also as the space of the possible hypothesis may be too vast to explore exhaustively. Alternative methods can be used in these cases, such as statistical sampling (Markov chain Monte Carlo).

The Bayesian paradigm can be used to build *Bayesian Networks*. A Bayesian network is a probabilistic graphical model, where nodes correspond to random variables and are connected with directed edges representing the conditional probabilities between the variables. The nodes are either observed random variables or latent/hidden random variables. Each node takes the parent(s) value(s) as input for computing the pdf of that node.

The graphical structure $\mathcal{M}$ of a Bayesian network has its nodes represented with $\mathbf{V}$ and the directed edges with $\mathcal{E}$. The graph structure is defined by the set of edges connecting the vertices, $\mathcal{M} = (\mathbf{V}, \mathcal{E})$. If an edge connects two nodes A and B and the arrow on the directed edge is towards node B, then A is referred to as the *parent* of B while B is referred to as the *child* of A. Figure 1.14 shows a Bayesian network diagram. There are 5 nodes representing random variables, $\mathbf{V} = A, B, C, D, E$, and the set of edges $\mathcal{E} = (A,B), (A,C), (B,D), (C,D), (D,E)$. Node A shows that a parent can have multiple children, and node D how multiple parents are possible. If there were no directions (arrows) on the edges then this graph would have a cycle between nodes A/B/C/D, but by following the edges a cycle is restricted. For a cycle to exist, the edges must point in the same direction until the original node is arrived at. This Bayesian network's joint probability $P(A,B,C,D,E)$ is equal to the product of the conditional probability relationships,

$$P(A,B,C,D,E) = P(A)P(B|A)P(C|A)P(D|B,C)P(E|D). \tag{1.64}$$

The factorisation for Bayesian networks from their joint probability into the product of conditional probability relationships can be done by considering the set of

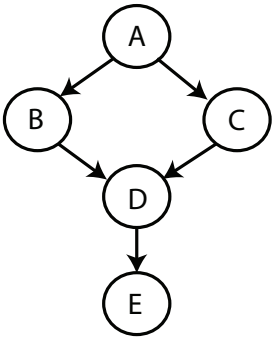


Figure 1.14: Bayesian network example

The circles are nodes representing random variables, which can be latent/hidden. The edges connecting the nodes show conditional independence relationships. At the end of the edges are arrows distinguishing on which side is a *parent* node (where the edge originates from), and a *child* node indicated by the arrow head. A parent node may have more than 1 children dependent on its value, shown by node A, and a node may have more than 1 parent shown by node D. The joint probability of the network, $P(A,B,C,D,E)$, is factorised into a product of the individual conditional probability relationships; $P(A)P(B|A)P(C|A)P(D|B,C)P(E|D)$.

random variables involved in the network; $X_1, X_2, \ldots, X_n$ identified with the index i which takes the values $1, \ldots, n$. $pa[i]$ is used to address the parent nodes of a node i in the network, and for the random variables which are the parents of X_i, $\mathcal{X}_{pa[i]}$ is used. The joint probability of the set of random variables in the graph can be written as a factorisation in terms of the parent random variables present in the joint probability;

$$P(X_1, X_2, \ldots, X_n) = \Pi_{i=1}^{N} P(X_i | \mathcal{X}_{pa[i]}). \tag{1.65}$$

The property of the *Markov blanket* can be used to find the factorisations of the random variables based on the conditional probability distributions. The Markov blanket for a particular node is the set of children, parents, and coparents (the other parents of this node's children). Given the Markov blanket for a particular node the rest of the nodes in the network are independent given this set. Using i again as an index for a certain node in the network, and $MB[i]$ as the set of nodes in the Markov blanket of node i, $\mathcal{X}_{MB[i]}$ are the random variables corresponding to $MB[i]$. The factorisation of eq 1.65 is denoted according to the Markov blanket,

$$P(X_k | X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n) = P(X_k | \mathcal{X}_{MB[i]}). \tag{1.66}$$

The conditional probabilities defined by the structure $\mathcal{M}$ of the Bayesian network can take on various functional forms. The probabilities can be from a Normal distribution, Gamma distribution, Poisson distribution and many others. These distributions have parameters which need to be defined. In the example of figure 1.14 the Markov blanket of node D is B, C, and E.

1.9 Hidden Markov models

The hidden Markov model (HMM) is a type of Bayesian network with hidden nodes and non-hidden nodes (observables). This model has been applied to natural language processing (NLP), speech recognition, and bioinformatics among other areas. For the problem of detecting the topology along the sites of the DNA sequence alignment which is not known beforehand, the complete sequence of data provides evidence towards the estimation of the topology at each site rather than utilising each single column of DNA independently (which would result in over-fitting).

The Figure 1.15 shows the structure of the hidden Markov model. The backbone of the chain connects the hidden state variables, S_t, where S_t is dependent on S_{t-1} and S_{t+1} from the Markov blanket. Each observation y_t is dependent on the parent hidden state variable S_t. The directed arrows indicate the parent to children relationship. The conditional independencies between nodes on this structure is given through the Markov blanket and is explained in Heckerman (1999). The probability of a single observation is independent of all the other observations and hidden state variables given its parent state variable,

$$P(y_t|y_1,\ldots,y_{t-1},y_{t+1},\ldots,y_N,S_1,\ldots,S_N) = P(y_t|S_t). \tag{1.67}$$

The state transitions in the sequence of observations are only dependent on the previous state variables,

$$P(S_{t+1}|S_1,\ldots,S_t,y_1,\ldots,y_t) = P(S_{t+1}|S_t). \tag{1.68}$$

For the given sequence of observations, $y_1,\ldots,y_N$ and state sequence $\mathbf{S}$ the joint probability can be found,

$$P(y_1,\ldots,y_N,S_1,\ldots,S_N) = \prod_{t=1}^{N} P(\mathbf{y}_t|S_t)P(S_t|S_{t-1})P(S_1). \tag{1.69}$$

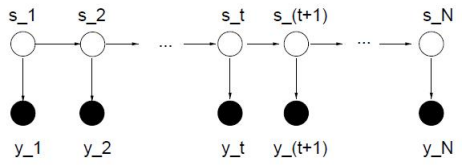


Figure 1.15: HMM structure

The hidden Markov model offers useful conditional independencies. The observations are denoted with y_t for the index values running along 1 to N for the sequence of observations, and each observation is associated with a hidden state variable S. The arrows show the association of parents to children. The individual observations can be found independently of other observations given the state variable value, $P(y_t|y_1,\ldots,y_{t-1},y_{t+1},\ldots,y_N,S_1,\ldots,S_N) = P(y_t|S_t)$, and the state variable transitions by $P(S_{t+1}|S_t)$. For a sequence of state transitions, $\mathbf{S}$, and observations the likelihood is the product of the observations and state transitions, $\prod_{t=1}^{N} P(y_t|S_t)\prod_{t=2}^{N} P(S_t|S_{t-1})$. Figure taken from Husmeier *et al.* (2005a).

The structure of the dependencies between the nodes is Markovian. The simplification allows the complexity of an exhaustive search of the hidden state sequences (eg. topologies or ratestates) to not be exponential in the number of hidden nodes (sequence length). With there being 3 possible topologies for a 4 strain alignment, the exhaustive search would have 3^N paths to search.

Given that there may be many states in the HMM, eq 1.69 may include the product (multiplicative) of many subsequently low numbers causing there to be underflow in the floating point arithmetic of the software used for implementing the model. To overcome this potential pitfall, the log likelihood is used. In later sections this may not be explicitly mentioned but it is used without mention to avoid inaccuracies.

Important algorithms for performing inference on HMMs are the Forward Algorithm in appendix A.7 which does *filtering*, the Forward-Backward Algorithm in appendix A.8 that does *smoothing* and the Viterbi Algorithm in appendix A.6 which performs *decoding*. Figure 1.16 illustrates these two processes.

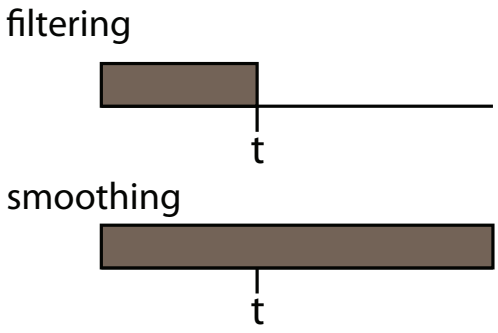


Figure 1.16: HMM filtering and smoothing

The figure illustrates the data used when conditioning on a particular site t when the process of *filtering* and *smoothing* is performed. The grey box depicts the amount of data used and the index t is used to show the place of the index in the data. Filtering is shown on the top of the two diagrams which is performed by the forward algorithm described in appendix A.7. The bottom diagram shows HMM smoothing performed by the forward-backward algorithm shown in appendix A.8.

1.9.1 Application of HMMs for inferring the topology states along sequence alignments

Here the application of HMMs in phylogenetics for inferring recombination break points is demonstrated and explained. McGuire *et al.* (2000) designed a likelihood method shown in figure 1.17. This figure shows a short DNA sequence alignment and indicates the fourth column. As each column is independent of other columns, the probability of the set of nucleotides at each site can be found. Not all the parameters used are presented here, such as the rate heterogeneity so that the motivation for HMMs used in detecting recombination is clearer. The set of candidate topologies for the given sequence alignment is shown in the left bottom. For the column of nucleotides the probability is found via $P(\mathbf{y}_t|S_t,\mathbf{w},\boldsymbol{\theta})$. The distribution of the topologies for each site is required when inferring the mosaic structure (break points) along the alignment $\mathbf{S}$.

The parameter for the probability of changing topologies along the sites is important for detecting recombination. The recombination state transition parameter's use is shown in figure 1.18 which has two subfigures. Subfigure a) illustrates the effect of the topology transition parameter which is expressed as the probability of not changing topology from one site to another site along the sequence alignment, $P(S_t|S_{t-1}) = \nu_S$ when $S_t = S_{t-1}$. The probability $1-\nu_S$ is

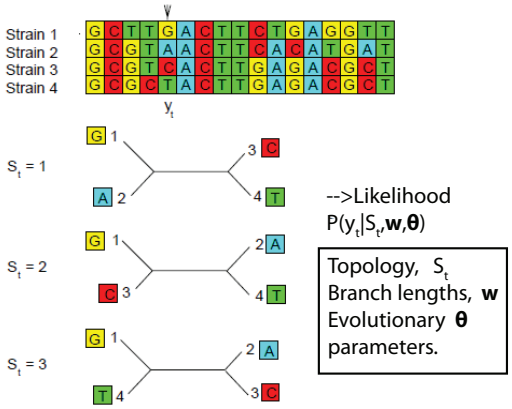


Figure 1.17: Sequence Alignment column parameters and likelihood

In the illustration, at the top there is a depiction of a sequence alignment, and the arrow above the fourth column indicates that the nucleotides at that site are to being examined. Each column of nucleotides is taken to be independent of each other and for this reason the model of the HMM is appropriate. $\mathbf{y}_t$ is used to represent the column of nucleotides, and the probability of the columns at site t in the alignment given the parameters of the topology S_t, branch lengths $\mathbf{w}$, and evolutionary parameters from the substitution model θ can be found; $P(\mathbf{y}_t|S_t, \mathbf{w}, \theta)$. Other essential parameters used in the model for calculating the probability of a column are introduced later. The 3 candidate topologies for the 4 strain sequence alignment are shown in the lower left. Obtaining the distribution of the topologies per site allows for inferring the sequence of the topologies along the whole alignment $\mathbf{S}$ as is described later. Figure taken from Husmeier *et al.* (2005a).

the probability of changing topologies and is uniformly distributed amongst the 2 possible candidates, $^{(1-\nu_S)}/_2$. Subfigure b) shows the effect of ν_R, the ratestate transition parameter. The ratestate parameter is the probability of remaining in the same ratestate which scales uniformly the complete set of branch lengths for the phylogenetic tree. The probability of transitioning into a different ratestate along the sequence is $^{1-\nu_R}/_{\tilde{K}}$, where the denominator is $\tilde{K}$. The number of ratestates is not restricted or known *apriori* unlike the number of topologies which is restricted by the number of sequence alignments.

The data, $\mathcal{D}$ of the sequence alignment is the set of the columns at each site $\mathcal{D} = (y_1, y_2, \ldots, y_N)$. The sequence of hidden topology states for the set of sites along the alignment is $\mathbf{S} = (S_1, S_2, \ldots, S_N)$. Eq 1.69 displays the factorisation of the HMM with only the dependency of the topology states. The HMM used, developed and tested uses different extensions. Developments further in the thesis involve including the parameter vector for the evolutionary nucleotide substitution parameters θ, the branch lengths $\mathbf{w}$, the ratefactor vector ρ, and the vector of relative codon rate heterogeneity λ. The HMM factorisation will be presented again with the introduction of the new parameters. Modelling the rate heterogeneity requires the factorial HMM that is discussed later in section 1.9.4. Introducing only the branch lengths and nucleotide substitution parameters without the ratefactors is;

$$\begin{aligned} P(\mathcal{D}, \mathbf{S}, \mathbf{w}, \theta, \nu_S) &= P(\mathbf{y}_1, \ldots, \mathbf{y}_N, S_1, \ldots, S_N, \mathbf{w}, \theta, \nu_S) \\ &= \prod_{t=1}^{N} P(\mathbf{y}_t | S_t, \mathbf{w}, \theta) \prod_{t=2}^{N} P(S_t | S_{t-1}, \nu_S) P(S_1) P(\mathbf{w}) P(\nu_S) P(\theta). \end{aligned} \tag{1.70}$$

The prior probabilities $P(\mathbf{w})$, $P(\nu_S)$, and $P(\theta)$ are assumed to be independent and the product of the three can be taken from the joint $P(\nu_S, \mathbf{w}, \theta) = P(\mathbf{w})P(\nu_S)P(\theta)$. The transition probability between states ($P(S_t|S_{t-1}, \nu_S)$) is dependent on the recombination parameter/probability, ν_S. For this model without rate heterogeneity being considered, the emission probabilities are:

$$P(\mathbf{y}_t | S_t, \mathbf{w}, \theta). \tag{1.71}$$

With the Kimura model, subsection 1.3.1, the nucleotide substitution parameters θ is the transition-transversion ratio t_s/t_v, and for the more complex model HKY in appendix A.3, $\theta = (t_s/t_v, \pi)$ (where π is defined in eq 1.18).

Because the different phylogenetic trees cannot share the same set of branch lengths, the branch lengths are a separate vector for each topology. The same

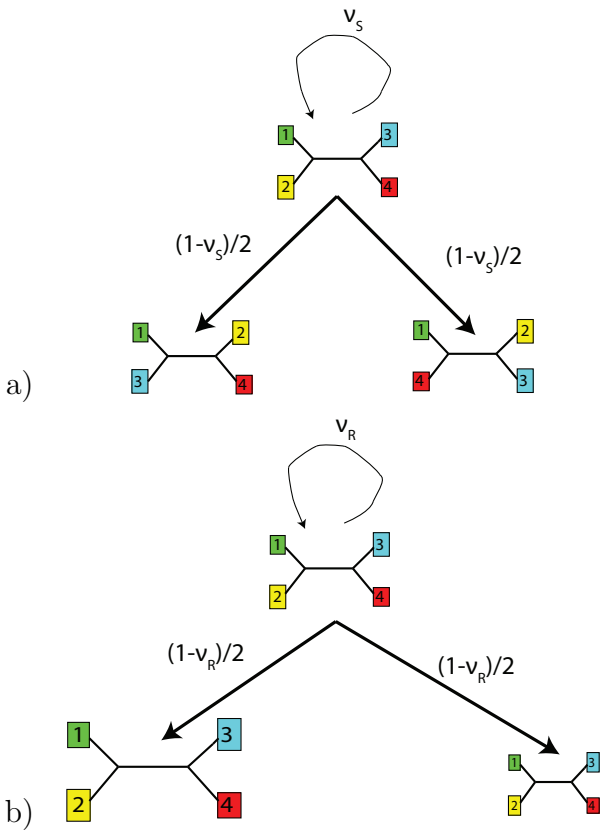


Figure 1.18: Transition Probabilities and Modelling Recombination

Subfigure a) shows how the transition probability ν_S works for changing the topology. ν_S is the probability of the topology not changing from one state to the other along the sites of the sequence alignment, $P(S_t|S_{t-1})$ where $S_t = S_{t-1}$. In a 4 strain DNA sequence alignment there are 2 other candidates and the probability $1-\nu_S$ is the probability of there being a topology change. The probability $1-\nu_S$ is distributed uniformly across the possible candidates giving a probability of $^{(1-\nu_S)}/_2$ for a transition to a different topology. Subfigure b) shows the analogous role for the transition parameter ν_R. This parameter is the probability of the rate state not changing between sites in the HMM model along the alignment $P(R_t|R_{t-1})$. The probability for changing a rate state is $1-\nu_R$ and is distributed uniformly amongst the possible rate states. The number of ratestates is not restricted to the number of strains in the alignment. There can be an arbitrary number of ratestates, so the transition probability into a particular ratestate is $^{1-\nu_R}/_{\tilde{K}}$, where $\tilde{K}$ is the number of rate components available to transition into. Figures adapted from Husmeier *et al.* (2005a).

holds for nucleotide substitution parameters and the relative ratefactors for the codon positions, but not for the ratefactors. It would be more accurate when writing the emission probability to have the subscripts shown:

$$P(\mathbf{y}_t|S_t,\mathbf{w}_{S_t},\theta_{S_t}). \tag{1.72}$$

To simplify the notation the subscripts conditioning on the relevant topology at the site t and the topology allocated, have been removed. The accumulation of the vectors for all the topologies are considered part of the denoted vectors; eg. $\mathbf{w} = (\mathbf{w}_1,\ldots,\mathbf{w}_K)$. The experiments and simulation performed were only on sequences of 4 species so the variable K is 3 at all times (in later stages K is used for different quantities).

Figure 1.19 shows the dependency structure of the HMM chain for the topology states at the sites along the DNA sequence alignment using the constructions of figure 1.17 and figure 1.18. t indexes a site in the alignment, the white nodes are the hidden state for the topology, the black nodes are the observations of the columns of nucleotides, $\mathbf{w}$ is the vector of branch lengths, and ν_S the topology state transition parameter for the probability of not changing topologies between sites. The arrows show the dependency structure between the nodes. The probability of the observations depends on the topologies and branch lengths but the model introduces other parameters as well such as the ratefactor value at each site and the factor accounting for the codon rate heterogeneity as well. The nucleotide substitution vector is also present in the model but not in the figure.

The topology state sequence $\mathbf{S}$ is a product of the probabilities of state transitions which can be homogeneous or heterogeneous state transitions. The sequence of topology state transitions is given by:

$$P(\mathbf{S}) = P(S_1,\ldots,S_N) = \prod_{t=2}^{N} P(S_t|S_{t-1})P(S_1) \tag{1.73}$$

where the probability of the first state $P(S_1)$ is set to be uniform over the set of possible values that it can take. The component $P(S_t|S_{t-1})$ which represents the hidden topology state transition probability as,

$$P(S_t|S_{t-1},\nu_S) = \nu_S^{\delta(S_t,S_{t-1})}\left(\frac{1-\nu_S}{K-1}\right)^{[1-\delta(S_t,S_{t-1})]}. \tag{1.74}$$

The function denoted by $\delta(.)$ is the Kronecker delta symbol taking the value of 1 when the transition is homogeneous $S_t = S_{t-1}$ and the value of 0 when the states are not equal (a recombination event).

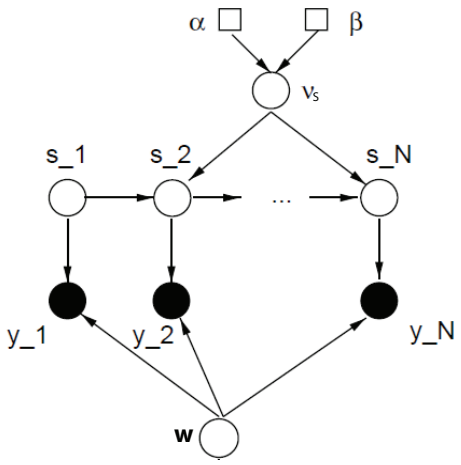


Figure 1.19: Diagram of the ν_{state} parameter for the detection of recombination. Topology state break points are modelled as changes in the hidden state variables along the topology HMM chain. The subscript t in the state variables of y and S denote the sites in the DNA sequence alignment. The black nodes, y, are observations which are not hidden ($\mathcal{D}$). The white nodes are the hidden states which must take the value of a particular candidate topology (of which there are 3 for 4 sequences in the alignment). The arrows between the nodes represent the conditional dependencies. α and β in squares are the hyper parameters for the beta distribution of ν_S. The probability (emission probability) of the observation at y_t is shown to be dependent on the topology state S_t and the vector of branch lengths $\mathbf{w}$. The emission probability is dependent, in the model used, on the vector of nucleotide substitution parameters θ and the ratefactor allocated to site t. Later in further sections the codon structure of DNA will also be an important variable for the probability emission at each site. The topology at a site t depends on the topologies at adjacent sites, S_{t-1} and S_{t+1} and the topology state transition parameter ν_S. Figure adapted from Husmeier *et al.* (2005a).

The state transition parameters are unknown beforehand, especially as the mosaic structure of the sequence alignment is also unknown. In many cases it is optimised using the Baum-Welch algorithm which contains an optimisation criteria of the likelihood of the sequence. In this work this approach is not taken, but a specific variant of the expectation-maximisation algorithm (EM algorithm) is used instead. Subsection 1.9.3 describes the algorithm and is presented in Husmeier and McGuire (2003). This implementation samples the value of the state transition parameters from the posterior distribution.

1.9.2 Beta distribution for the transition parameter between the hidden states in the hidden Markov model

In the appendix section A.4 introduces the beta distribution. The beta distribution is used as a prior for the HMM hidden state transition parameter ν. It is a conjugate to the binomial distribution, meaning that a Beta distribution remains when the prior combined with the binomial. The recombination parameter, ν_S [1], is a binomial random variable. When the data set is large the effect of the prior becomes less significant against the likelihood of the data. (these alpha and beta parameters for the Beta distribution should not be confused with the alpha and beta values of the forward-backward algorithm)

From figure 1.20, subfigure a) shows a set of subplots of the Beta distribution for different values of its two parameters α and β. For each plot shown the Beta distribution used had the parameter $\beta = 2$. The α parameter value is changed so that the mean of the distribution would be equal to 0.5, 0.6, 0.7, 0.8, 0.9, and 0.95. It is visible how the density shifts towards the mean and is constrained between the domain values of $[0, 1]$.

The subfigure b) from figure 1.20 shows the transition probability between homogeneous and non-homogeneous regions. Each change of state requires the probability $1-\nu$, so an unbroken single stretch of a particular state value would push the ν value very close to 1 maximising the log-likelihood of the model for data. Frequent state changes would lower the value of ν to reduce the penalisation of $1-\nu$ for the large number of times it will occur. Given the value of ν the expected segment length n can be computed. The probability for the segment

[1] breakpoint parameter for the topology states

length n is,

$$P(n) = \nu^{n-1}(1-\nu). \tag{1.75}$$

The expected segment length (average n for $N \to \infty$) is,

$$\langle n \rangle = \Sigma_{n=1}^{N} nP(n) = (1-\nu)\Sigma_{n=1}^{N} n\nu^{n-1} = (1-\nu)\frac{d}{dv}\Sigma_{n=1}^{N}\nu^{n} = (1-\nu)\frac{d}{dv}\frac{1}{1-\nu} = \frac{1}{1-\nu}. \tag{1.76}$$

1.9.3 Sampling from the posterior distribution of the hidden state transition probabilities

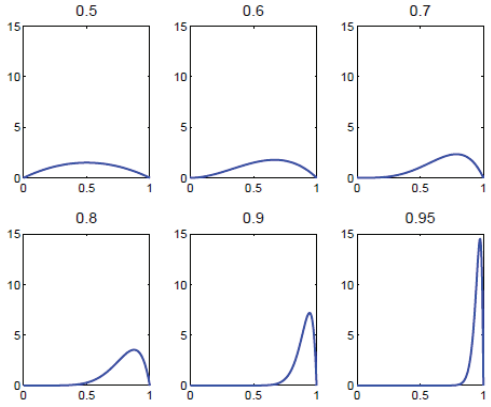
The approach used in this thesis to sample the state transition parameters for the HMM differs significantly from the Baum-Welch algorithm. An optimisation criteria is not used but a sampling method is used which draws samples from the posterior distribution. Another important aspect is the simplification in the modelling approach taken. It is assumed that the probability of the transitions is identical in transitions between non-identical states. Figure 1.18 depicts the modelling approach of the hidden state transitions with a diagram. In terms of the transition matrix A, $A_{k,l} = A_{k,m}$ for these terms which are off the diagonal, and are equal to the probability of changing states $(1-\nu)$ divided by the number of possible non-homogeneous state transitions; $A_{k,l} =^{(} 1-\nu_S)/_{K-1}$. For the topology state transitions ν_S is used, and ν_R is used for the ratefactor state transitions. The diagonal entries $A_{k,k}$ (probabilities of not changing state) are ν_S for the topology transition probabilities and ν_R for the ratefactor state changes.

This approach is described in Husmeier and McGuire (2003) on page 319 and 320. This transition probability is a recombination parameter because the state transitions represent a change point in the topology state along the alignment (for ν_S). The state transition parameter ν is a binomial random variable, and the beta distribution is conjugate to it (discussed in subsection 1.9.2). The sampling of the transition probability is from the pdf of the beta distribution in eq A.11 and using the appropriate symbol of ν (either ν_S or ν_R),

$$P(\nu) = \text{Beta}(\nu|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\nu^{\alpha-1}(1-\nu)^{\beta-1}. \tag{1.77}$$

The parameters α and β change the value of the mean and the variance of the distribution. They are used in the prior belief of the number of change point observations made along the sequence of hidden states. The Kronecker delta

a)



b)

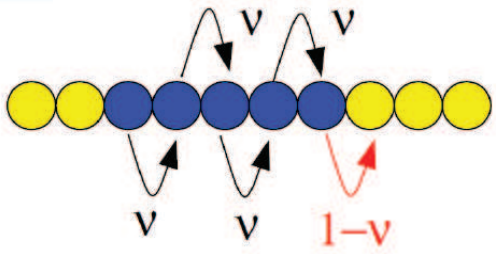


Figure 1.20: Illustrations of the Beta Distribution for the hidden state transition parameter

The transition probability between the hidden states (ν) of the HMM, for the topology states it is ν_S and for the ratestates ν_ρ. Subfigure a) shows a set of subplots of the probability density across the domain of the Beta distribution; $[0, 1]$. The parameter $\beta = 2$ for each plot, the value of the mean is equal to the value displayed above the plot, $\mu = {}^{\alpha}/_{(\alpha+\beta)}$ from which the value of α can be inferred. The α and β parameters can be interpreted as the number of observations seen of the two possible outcomes of the Bernoulli trial, and are appropriate for the modelling of 'state change/no state change'. Subfigure b) shows the effect of the compound probability for the ν. Longer stretches of sites (homogeneous topology regions) favour a high transition probability to increase the product of the probability for the log likelihood of the model. Shorter homogeneous regions favour lower values of ν to reduce the penalisation of more frequent state changes. From the value of ν the expression for the expected number of sites a homogeneous region will contain; $\langle n \rangle = {}^{1}/_{(1-\nu)}$. Figures taken from Husmeier *et al.* (2005a).

function is used($\dot{\delta}()$) for $\delta(S_t, S_{t-1})$ to equal 1 when $S_t = S_{t-1}$ and 0 when $S_t \neq S_{t-1}$. A variable Ψ is defined to represent the number of homogeneous state changes along the hidden state trajectory, $\Psi = \sum_{t=1}^{N-1} \delta(S_t, St+1)$. The joint of the model is directly proportional to the prior times the likelihood of the state transition probability,

$$P(\mathcal{D}, \mathbf{S}, \mathbf{w}, \theta, \nu_S) \propto \nu_S^{\Psi+\alpha-1}(1-\nu_S)^{N-\Psi+\beta-2}, \tag{1.78}$$

which is required for eq 1.105. This proportionality can be seen from eq 1.87. The normalisation factors from the resulting beta distribution were omitted and the normalised expression can be used to give the posterior distribution of the transition probabilities to sample from,

$$P(\nu_S|\mathcal{D}, \mathbf{S}, \mathbf{w}, \theta) = \mathrm{B}(\nu_S|\Psi+\alpha, N-1-\Psi+\beta). \tag{1.79}$$

With the samples of the transition probability the A matrix can be made and the trellis of the HMM for the state sequences can be calculated.

1.9.4 The phylogenetic factorial hidden Markov model (phylo-FHMM)

Subsection 1.9 presented the hidden Markov model (HMM) and the equations which defined it. HMMs were originally first used in the field of phylogenetics in Felsenstein and Churchill (1996) and Yang (1995) for modelling rate variation among the sites in the sequence alignment. A set of finite rates were chosen to be applied along the alignment and the HMM provided correlations between the rates of the neighbouring sites as for the topologies previously described. It is assumed in these papers that the topology (phylogenetic tree) is known beforehand or that another method has inferred it. Subsection 1.7 introduces the evolutionary aspect of rate heterogeneity and gives a brief overview of its biological importance. This subsection presents the work of extending the HMM to a factorial hidden Markov model (FHMM) that will combine the estimation of topology break points and rate variation break points in one model. The subsection 1.7 serves as a primer for this concept.

The model applying the HMM to topologies was presented in Husmeier and Wright (2001) and prior to that in McGuire *et al.* (2000) for the purpose of detecting recombination in DNA sequence alignments as described in subsection 1.9.1. The dependency between the sites is restricted to the Markov blanket explained

in Heckerman (1999). For a given DNA sequence alignment of N columns, a HMM of N hidden states is made, (1 hidden state to represent a topology at a given site/column) and the index t takes values within the range of the sequence $1 \leq t \leq N$. Each hidden state for the HMM is denoted by S, the state at a particular site is S_t, and the topology that the hidden states can take on at each site of the HMM chain is $S_t \in \tau_1, \ldots, \tau_K$. The symbol K is used to denote the total number of topologies. The total number of topologies is determined by the number of sequences present in the DNA sequence alignment, m, which is given by the formula presented in eq 1.2. The observed states are the nucleotides present in the alignment where all the sites are grouped under $\mathcal{D}$. At each column is a vector of nucleotides, $\mathbf{y}_t$ whose emission probability is dependent on the topology, as defined in eq 1.71, the branch lengths $\mathbf{w}$, and the parameters for the nucleotide substitution model $\boldsymbol{\theta}$. As described in the publication of the FHMM in Husmeier (2005), the nucleotide substitution parameters can be excluded from the sampling procedure. This creates a reduction in the computational costs, and has a small effect on the topology statesequences sampled.

The phylogenetic factorial hidden Markov model (phylo-FHMM) is presented in Husmeier (2005). The factorial hidden Markov model consists of two *apriori* independent HMM chains. The HMM attempting to infer a correct topology sequence along an alignment belongs to one chain and the ratefactors in the other. If there are large variations in the rate of mutation (the complete set of branch lengths changes their magnitude), then the model is no longer susceptible to misinterpret these changes as topology break points. As a result rate variation mosaics can be inferred correctly. The FHMM can distinguish between recombination events and rate variation along a sequence alignment.

Each HMM chain is combined into the factorial HMM (FHMM). Where changes between topologies represent recombination events the changes in the scalings represent different states of selective pressure upon the organism's genome. The algorithms for inference on the HMM still apply to the individual chains whose parameters both will affect the global posterior probability of the complete model (forward-backward algorithm A.8, nested Gibbs sampling A.9, and the stochastic forward backward algorithm 1.9.5). The inference is performed in a hierarchical Bayesian model. Groups of the parameters in the model are made and then Gibbs sampling is performed upon the groups of parameters. Figure 1.21 shows the Bayesian network of the FHMM. The hidden topology states S_t along

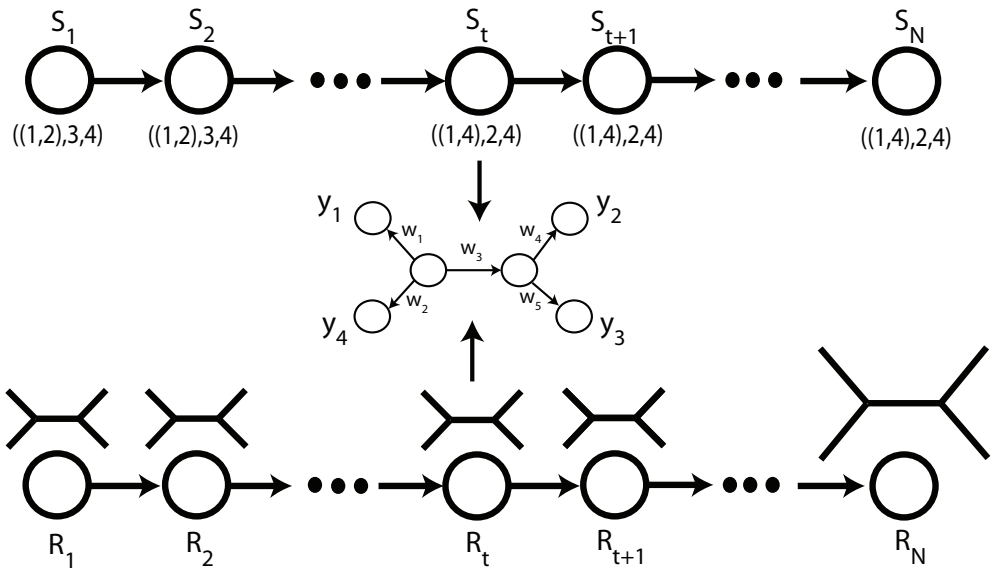


Figure 1.21: The phylogenetic FHMM draw as a Bayesian network.
The figure shows the factorial hidden Markov model (FHMM) applied to phylogenetic trees. There are two HMM chains, 1 for the hidden topology states and the second for the hidden ratefactor states. At each site t in the DNA sequence alignment there is a hidden state drawn as an empty circle along the HMM chains. The ratefactors at each site represent a scaling parameter for the vector of normalised branch lengths, $\mathbf{w} = \dot{\mathbf{w}} \times \rho$. The topologies are presented as a branching order on the bifurcations and which strains are adjacent to each other. At the site t in the alignment the effect of the two chains on the resulting phylogenetic tree is displayed.

the alignment are shown on the top chain and the bottom chain for the ratestates R_t taking the value of a ratefactor ρ. The branch lengths at each site in the alignment are a result of the scaling of the normalised vector of branch lengths by the ratefactor selected at that site along the HMM chain; $\mathbf{w} = \dot{\mathbf{w}} \times \rho$.

The number of ratefactors available $\tilde{K}$ are predefined and their values fixed. Further sections describe the extension to relax these constraints. Each ratestate R can take on the value of the ratefactors $\rho_i \in \rho_1, \ldots, \rho_{\tilde{K}}$. The ratefactor vector is denoted by ρ.

In the work of Husmeier (2005) an independent product of exponential distributions is put on the branch lengths. This was introduced in Suchard *et al.* (2003). The product of the independent branch lengths for the branch length

vector $\mathbf{w}$ is applied at each site,

$$P(\mathbf{w}|\rho) = \prod_i P(w_i|\rho), \tag{1.80}$$

and the exponential distribution for each branch length is

$$P(w_i|\rho) = \frac{1}{\rho} e^{\frac{-w_i}{\rho}}. \tag{1.81}$$

The model follows no-common-mechanism model (NCM) of Tuffley and Steel (1997) where each of these branch length vectors is applied independently at each site in the alignment (contrast to assuming a branch length vector for the complete alignment). The integration of the branch lengths are analytically tractable since the likelihood is conjugate to this prior,

$$P(\mathbf{y}_t|S_t,\rho_t) = \int P(\mathbf{y}_t|S_t,\mathbf{w})P(\mathbf{w}|\rho_t)d\mathbf{w}. \tag{1.82}$$

The ratefactor without the individual branch lengths along the phylogenetic tree acts as an average amount of evolutionary change representing the average number of mutations, or average branch length, which is sampled along the second HMM chain in the FHMM.

Eq 1.73 and eq 1.74 show the equations for the topology states that mirror these for the rate state transitions along the HMM chain. The ratestate sequence $\mathbf{R}$ has its dependency modelled for the whole sequence as the topology states when considering the probability of the state transitions,

$$P(\mathbf{R}) = P(R_1,\ldots,R_N) = \prod_{t=2}^{N} P(R_t|R_{t-1})P(R_1). \tag{1.83}$$

The probability of the first state $P(R_1)$ is set to be uniform over the set of possible values that it can take. The component $P(R_t|R_{t-1})$ which models the hidden ratestate transition probability is,

$$P(R_t|R_{t-1},\nu_R) = \nu_R^{\delta(R_t,R_{t-1})} \left(\frac{1-\nu_R}{\tilde{K}-1}\right)^{[1-\delta(R_t,R_{t-1})]}. \tag{1.84}$$

The prior on ν_R follows a beta distribution mirroring that for the topology state transition prior.

The emission probabilities defined previously are applied to the HMMs and values of the hidden topology states. The branch lengths were taken into consideration as well. Given the option of choosing the independent product of

exponential priors on the branch lengths from Suchard *et al.* (2003) with the no-common-mechanism model, the branch lengths can be omitted and the two emission probabilities are shown for both cases. The form of the dependency of the column data when the branch lengths are removed from the model is,

$$P(\mathbf{y}_t|S_t,R_t,\theta), \qquad (1.85)$$
$$P(\mathbf{y}_t|S_t,R_t,\theta_{S_t}).$$

Then the form where they are included,

$$P(\mathbf{y}_t|S_t,R_t,\mathbf{w},\theta) \qquad (1.86)$$
$$P(\mathbf{y}_t|S_t,R_t,\mathbf{w}_{S_t},\theta_{S_t}),$$

which are extended from equations 1.71 and 1.72 by use of the FHMM allowing the columns of nucleotides to be dependent on both the topology and ratefactor at a site.

Extending eq 1.70 of the likelihood of the HMM, to account for the HMM of ratestates has this form (excluding the branch lengths),

$$P(\mathcal{D},\mathbf{S},\mathbf{R},\nu_S,\nu_R) = \qquad (1.87)$$
$$= \prod_{t=1}^{N} P(\mathbf{y}_t|S_t,R_t) \prod_{t=2}^{N} P(S_t|S_{t-1},\nu_S) \prod_{t=2}^{N} P(R_t|R_{t-1},\nu_R) \times P(S_1)P(R_1)P(\nu_S)P(\nu_\rho).$$

Using the marginal posterior probabilities of the ratestates along the sites of the sequence alignment will allow comparisons of the probabilities of the various possible ratestates. Mosaic structures along the alignment allude to differences in selective pressure over the genome.

Here the method used is stochastic forward backward algorithm which samples the state sequences of the HMM chain and is described in section 1.9.5 (the same applies for both the topology and rate states);

$$P(R_t|R_{t+1},\ldots,R_N,\mathbf{y}_1,\ldots,\mathbf{y}_N) = \frac{P(R_{t+1}|R_t=k)\alpha_t(R_t=k)}{\Sigma_i P(R_{t+1}|R_t=i)\alpha_t(R_t=i)}. \qquad (1.88)$$

The complete state sequences for the rate or topology states is obtained by marginalising over the joint posterior distribution of $P(\mathbf{S},\mathbf{R},\nu_S,\nu_R|\mathcal{D})$ which is directly proportional to the joint likelihood of the model. The complete state sequences for the ratefactor and topologies can be obtained from these equations:

$$P(\mathbf{S}|\mathcal{D}) = \sum_R \int P(\mathbf{S},\mathbf{R},\nu_S,\nu_\mathbf{R}) d\nu_S d\nu_R \qquad (1.89)$$

$$P(\mathbf{R}|\mathcal{D}) = \sum_S \int P(\mathbf{S},\mathbf{R},\nu_S,\nu_\mathbf{R}) d\nu_S d\nu_R. \qquad (1.90)$$

These two integrals are analytically intractable and therefore are approximated by drawing statistically consistent samples using MCMC. Samples are drawn from the joint posterior distribution using a Gibbs sampling procedure (Casella and George (1992)) which is described in subsection 1.10.2. In this Gibbs sampling procedure one parameter group (eg. one of the hidden state transition parameters or one of the statesequences) is sampled after another and the sampler cycles through all the parameters iteratively. The subsection on Gibbs sampling presents this in more detail.

1.9.5 Stochastic forward-backward algorithm

In contrast to the objective of the Viterbi algorithm (Appendix A.6) in finding the mode of the distribution $P(S_1,\ldots,S_N|\mathbf{y}_1,\ldots,\mathbf{y}_N)$, the objective of the stochastic forward-backward algorithm is to sample (instead of optimising) a whole state sequence from the conditional distribution $P(S_1,\ldots,S_N|\mathbf{y}_1,\ldots,\mathbf{y}_N)$.

The discussion of this method and application to HMMs in detecting mosaic sequences, is given in Werhli *et al.* (2006). The stochastic forward-backward algorithm is a modification of the forward-backward algorithm described in Appendix A.8, and is used within an unnested Gibbs sampling procedure. The computational costs are reduced from those of the nested Gibbs-within-Gibbs sampling procedure described in Appendix A.9. The computational costs are reduced due to the improvement in the mixing and convergence of the Markov chain. The study of Werhli *et al.* (2006) has empirical results showing two orders of a magnitude reduction, from $10^5, 10^6$ to $10^3, 10^4$ in the steps of test simulations.

It is essential to utilise the structure of the HMM for computational efficiency. Brute force (naive) approaches require exponential computational times in terms of the length of the HMM chain. The stochastic forward backward algorithm samples the whole state sequence of the posterior distribution. The motivation of the algorithm is in the following sequence of equations:

$$P(S_t|S_{t+1},\ldots,S_N,\mathbf{y}_1,\ldots,\mathbf{y}_N) \tag{1.91}$$

$$\propto P(S_t,S_{t+1},\ldots,S_N,\mathbf{y}_1,\ldots,\mathbf{y}_N) \tag{1.92}$$

$$= P(\mathbf{y}_{t+1},\ldots,\mathbf{y}_N,S_{t+1},\ldots,S_N|S_t,\mathbf{y}_1,\ldots,\mathbf{y}_t)P(S_t,\mathbf{y}_1,\ldots,\mathbf{y}_t) \tag{1.93}$$

$$= P(\mathbf{y}_{t+1},\ldots,\mathbf{y}_N,S_{t+1},\ldots,S_N|S_t)\alpha_t(S_t). \tag{1.94}$$

Here α_t is the alpha parameter from the forward backward algorithm

$P(S_t, \mathbf{y}_1, \ldots, \mathbf{y}_t)$ described in eq A.28. Continuing from the above equations,

$$= P(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N, S_{t+2}, \ldots, S_N | S_{t+1}) P(S_{t+1} | S_t) \alpha_t(S_t) \tag{1.95}$$

$$\propto P(S_{t+1} | S_t) \alpha_t(S_t). \tag{1.96}$$

This last step can be done since $P(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N, S_{t+2}, \ldots, S_N | S_{t+1})$ is independent of S_t and cancels out in the normalisation as a constant factor. The alphas represent the set of forward probabilities (likelihoods under the model) with the first forward sweep of the forward-backward algorithm, and it is for the first t columns in the sequence alignment. At the last site in the alignment the distribution of the alphas (at site $t = N$), is for all the hidden states at that site given the complete set of observations made. The initialisation of the stochastic forward-backward algorithm is,

$$P(S_N = k | \mathbf{y}_1, \ldots, \mathbf{y}_N) = \frac{\alpha_N(S_N = k)}{\Sigma_i \alpha_N(S_N = i)}. \tag{1.97}$$

Using the last state in the chain as a starting point, a recursion can be made from $N - 1$ to 1 for all the states in the chain. For the consecutive state continuing until the first state at the first site can be found with,

$$P(S_t | S_{t+1}, \ldots, S_N, \mathbf{y}_1, \ldots, \mathbf{y}_N) = \frac{P(S_{t+1} | S_t = k) \alpha_t(S_t = k)}{\Sigma_i P(S_{t+1} | S_t = i) \alpha_t(S_t = i)}. \tag{1.98}$$

This way allows a state sequence to be sampled from the posterior distribution and in polynomial running time. It requires the forward algorithm to be run first.

This method improves on the Gibbs-within-Gibbs sampling of the hidden state sequences described in subsection A.9. The approach of sampling the whole state sequence with Gibbs-within-Gibbs sampling has lower computational costs than that of the stochastic forward backward algorithm, but it has poorer mixing and convergence; see Werhli *et al.* (2006) for an empirical comparison. The nested Gibbs-within-Gibbs scheme is described in Husmeier and McGuire (2003), where each state S_t is sampled in a separate step of a Gibbs sequence. Boys *et al.* (2000) describes the methods but also introduces this improved approach used here.

1.10 Sampling methods

In this section a set of sampling methods are described. These methods are used in the rest of the thesis.

1.10.1 Markov Chain Monte Carlo and Metropolis Hastings

Subsection 1.3.1 described some of the properties of Markov chains in the context of models of nucleotide substitution and some parts will overlap here. A Markov chain or Markov process has the Markov property that future states of the process depend only on the present state. The future state transitions are dependent only on the present state and not on past states. Stochastic processes with this property for their conditional probability distributions are called Markov processes. For discrete state values the property's effect on the conditional distribution can be seen with this equation:

$$P(X_n = x_n | X_{n-1} \dots X_0 = x_0) = P(X_n = x_n | X_{n-1} = x_{n-1}). \tag{1.99}$$

Here the X_i denote states and x_i the values the states take on.

A Markov chain is used to refer to a Markov process which has a discrete and finite set of state space variables and time being a discrete set as well. The probabilities between the state changes are called *transition probabilities*. The transition probabilities can be represented as the directed edges between nodes on a directed graph.

The Markov chains of concern here are *irreducible*, meaning that there is a non-zero probability of transition between any of the states in the Markov chain. These Markov chains are *aperiodic*, as they do not have a certain amount of time required before a state can be revisited. A state in the Markov chain is *recurrent* if there is a finite time for a specific state to be revisited. There is the property of *ergodicity* where a state i is *ergodic* if it is aperiodic and positive recurrent. For a Markov chain to be ergodic all of the states must be ergodic as well, and the model on the states defining the state transitions must allow any state to reach another via a finite number of steps. This is independent of the initial state as it holds for all the states. *Reversible* Markov chains are chains which are reversible if the equation of *detailed balance* is satisfied (where x is the initial state and x' the new proposed state):

$$P(x)Q(x,x') = P(x')Q(x',x). \tag{1.100}$$

Here the symbol $P(x)$ refers to the probability of being in a state x and $Q(x,x')$ the proposal probability of the state. If the transition probabilities do not change during the progress of the state transitions then the Markov chain is said to be *homogeneous*. The distribution of the states, independent of the initial distribution,

is the *stationary distribution* of the Markov chain (the stationary distribution is the probabilities used in the above eq 1.100 for satisfying detailed balance).

The Metropolis-Hastings algorithm Metropolis *et al.* (1953) is a Markov chain Monte Carlo (MCMC) method which simulates a Markov chain on a distribution of interest which cannot be sampled from directly. The samples are not independent from each other as in Importance sampling but have first order dependence as with the standard Markov chains. This allows for samples to be concentrated around regions of higher density that have larger influence on the results. Considering a distribution $P(\cdot)$, from which samples need to be drawn from, the state x with density $P(x)$, can have another state x' (with density $P(x')$) which is proposed from another distribution (a proposal distribution) with density $Q(x,x')$. The reverse proposal density is $Q(x',x)$, and the acceptance function that satisfies detailed balance in eq 1.100 is,

$$A_{\min} = \text{Min}\left(1, \frac{P(x')}{P(x)}\frac{Q(x',x)}{Q(x,x')}\right). \tag{1.101}$$

The minimum between the comma separated values is chosen as the probability for accepting the the move from x to x'.

From the property of ergodicity of the converged Markov chain whose samples were drawn with the Metropolis-Hastings algorithm, the stationary distribution will not change and gives a correct approximation to the underlying distribution of $P(\cdot)$. A description of how the Metropolis algorithm is used in each section is given. The convergence is possible due to the property of ergodicity of Markov chains that will not depend on their initial state as the number of samples increases.

1.10.2 Gibbs sampling

Gibbs sampling is a variant of the Metropolis Hasting MCMC algorithm discussed in subsection 1.10.1. The algorithm generates statistical samples consistent with the joint probability distribution of two or more random variables. As with MCMC, the goal is to sample in the limit of convergence from the joint distribution by generating samples that converge to the correct distribution. Gibbs sampling allows sampling from the conditional distributions of each variable in turn.

It is applicable in cases where the conditional distribution of the parameters can be computed. Each sample of a single parameter conditional on the rest

of the parameter vector is consistent with the distribution and this is done in a sampling scheme where each of the parameters is sampled in the same way. The scheme uses a burn-in and sampling phase, as with MCMC, and each iteration of the simulation sequentially samples each of the parameters one after the other. The stationary distribution of the desired joint distribution of the parameters is achieved from the ergodicity of the Markov chain. The proposal distribution satisfies detailed balance and the steps of picking a particular index (parameter) in the parameter vector, and sampling it conditional on the rest of the variables is explained in Neal (1993).

A component k is chosen from the vector $\mathbf{x}$ holding all of the variables (the order does not affect the correctness of the sampling procedure). The conditional distribution is $P(x_k^{i+1}|x_j : j \neq k)$, and the probability of a vector is $P(\mathbf{x}) = P(x_k^i|x_j : j \neq k)P(x_j : j \neq k)$. The MCMC acceptance function is

$$\mathrm{A}_k\left(x_k^i, x_k^{i+1}\right) =_{\min} \left[1, \frac{P(x_k^{i+1}|x_j^{i+1} : j \neq k)P(x_j^{i+1} : j \neq k)P(x_k^i|x_j^{i+1} : j \neq k)}{P(x_k^i|x_j^i : j \neq k)P(x_j^i : j \neq k)P(x_k^{i+1}|x_j^i : j \neq k)}\right] = 1. \tag{1.102}$$

The Gibbs sampling steps are therefore always accepted and the acceptance function is no longer needed. With each iteration each variable is sampled in turn.

Subsection 1.9.4 describes the FHMM model and does not include the description of the Gibbs sampling procedure for the parameters. The procedure in each iteration samples one parameter group conditional on the rest of the parameters. The symbol (i) denotes the iteration number in the simulation of the Markov chain and to obtain the $(i+1)$ iteration the following scheme is used:

$$\mathbf{S}^{(i+1)} \sim P(\cdot|\mathbf{R}^{(i)}, \nu_S^{(i)}, \nu_R^{(i)}, \mathcal{D}) \tag{1.103}$$

$$\mathbf{R}^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)}, \nu_S^{(i)}, \nu_R^{(i)}, \mathcal{D}) \tag{1.104}$$

$$\nu_S^{(i+1)} \sim P(\cdot|\mathbf{R}^{(i+1)}, \mathbf{S}^{(i+1)}, \nu_R^{(i)}, \mathcal{D}) \tag{1.105}$$

$$\nu_R^{(i+1)} \sim P(\cdot|\mathbf{R}^{(i+1)}, \mathbf{S}^{(i+1)}, \nu_S^{(i+1)}, \mathcal{D}). \tag{1.106}$$

The order of these equations can be changed as the samples are consistent with the posterior distribution. Equation 1.103 and equation 1.104 sample the hidden state trajectory of the a priori independent HMM chain via the stochastic forward-backward algorithm described in subsection 1.9.5. Equations 1.105 and 1.106 can be performed by sampling from beta distributions described in subsection 1.9.2.

1.10.3 Reversible Jump Markov chain Monte Carlo

Reversible jump Markov chain Monte Carlo (RJMCMC) is presented in the paper of Green (1995). It is an extension of MCMC described in subsection 1.10.1. The improvement made is that the dimensionality of the parameter vectors are sampled according to the posterior distribution rather than being of fixed length. This very useful extension is used in this work to facilitate the ability to model variable numbers of ratefactor components rather than fixed numbers of ratefactor components which would otherwise have been the case. The RJMCMC method can also be referred to as trans-dimensional MCMC.

For a parameter vector ρ, there is a length for the number of components given by $\tilde{K}$, and this is denoted by $\rho_{\tilde{K}}$. A new value of the dimensionality for $\tilde{K}'$ is denoted by; $\rho_{\tilde{K}'}$. The number of parameters can increase via a *birth* or *death* move:

$$(\rho, u) = (\rho_1, \rho_2)\,, \tag{1.107}$$

which increases or decreases the number of components by 1. The variable u used here is a sampled value from some distribution. The two parameters ρ_1 and ρ_2 are independent of each other. There are two other moves *split* and *merge* which are not used in this thesis, but are discussed in appendix subsection A.12. For the birth and death moves to be reversible, a bijective function is required between the parameter spaces proposed, which is shown in appendix section A.16.

The RJMCMC scheme requires the likelihood ratio, the prior ratio, the inverse proposal probability ratio, and the Jacobian to be defined. The form of the acceptance function is similar to that of Metropolis Hastings. The Jacobian is added to normalise the volume of the space when dimensions are added or removed. The Jacobian allows the sampler to continue to satisfy detailed balance, which is needed for the sampling scheme to remain statistically consistent with the model inference is being performed on. Additionally the Jacobian allows for potential parameter transformations that is described in more detail in the appendix referenced from the previous paragraph. The form for the acceptance equation for a RJMCMC move is,

$$\{1, \text{likelihood ratio} \times \text{prior ratio} \times \text{Mininverse proposal probability ratio} \times \text{Jacobian}\}. \tag{1.108}$$

The abbreviations used for these terms respectively will be LR, PR, and IPPR for the first 3 terms above.

The sampling scheme chooses between 3 different moves which are birth, death and relocation. In the relocation step the dimensionality of the parameter vector does not change and the values of the vector components of $\rho_{\tilde{K}}$ are sampled via MCMC. These can also be referred to as within-model moves and between-model moves.

The acceptance of a specific birth move will be a function of the minimum value between 1 and another term:

$$A_b = Min\{\left(\frac{P(\mathcal{D}|\rho_{\tilde{K}'})}{P(\mathcal{D}|\rho_{\tilde{K}})} \times \frac{P(\rho_{\tilde{K}'})}{P(\rho_{\tilde{K}})} \times \frac{P(death)P(\tilde{K}' \to \tilde{K})}{P(birth)Q(\rho')P(\tilde{K} \to \tilde{K}')} \times \text{Jacobian}\right)\}. \tag{1.109}$$

The death move is a reciprocal of the birth move acceptance term.

Chapter 2

Addressing intrinsic inconsistencies of various recent Bayesian methods for detecting recombination

Here is presented the work that lead to the journal publication Husmeier and Mantzaris (2008). The work is related to three recent Bayesian methods for detecting recombination in DNA sequence alignments: the multiple change-point model (MCP) of Suchard *et al.* (2003), the dual multiple change-point model (DMCP) of Minin *et al.* (2005), and the phylogenetic factorial hidden Markov model (PFHMM) of Husmeier (2005). The idea underlying the MCP is to segment the DNA sequence alignment by the insertion of change points, and to infer different phylogenetic trees and nucleotide substitution rates for the separate segments thus obtained. Inference is carried out in a Bayesian way. Of particular interest are the number and locations of the change points, which mark putative recombination breakpoints. Starting from a truncated Poisson prior, the number of change points is sampled from the posterior distribution with reversible jump (RJ) Markov chain Monte Carlo (MCMC). A disadvantage of this approach is the inability of the model to distinguish between recombination and rate heterogeneity. This shortcoming is addressed in the DMCP, where two separate change-point processes associated with the phylogenetic tree topology and the nucleotide substitution rate are employed. A related but different modelling paradigm is provided by the PFHMM, where two a priori independent hidden Markov chains are introduced, whose states represent the tree topology and nucleotide substitution rate, respectively. The three models described above have one feature in

common: different sites in the sequence alignment are associated with separate branch lengths, which allows the latter to be integrated out analytically. This is convenient, as the marginal likelihood of the tree topology, the nucleotide substitution rate, and further parameters of the nucleotide substitution model (like the transition- transversion ratio) can be computed in closed from. In this way, the computational complexity of sampling break points (MCP,DMCP) or hidden state sequences (PFHMM) from the posterior distribution with MCMC is substantially reduced. The subject of the present work is to investigate the effect of the approximation on which the analytic integration of the branch lengths is based. We will demonstrate that as a consequence of this approximation, the resulting model may predict spurious topology changes. A clearer analysis of the underlying approximation reveals that the resulting model exhibits a behaviour very similar to maximum parsimony, and that it is intrinsically susceptible to the systematic failure in the Felsenstein zone (Felsenstein, 1978b) and described in subsection 1.4.2. We propose a modification of the PFHMM without the aforementioned distributional approximation for the branch lengths. This modification increases the computational complexity of the inference scheme, as the branch lengths have now to be numerically sampled from the posterior distribution. However, we demonstrate that the resulting model will avoid the prediction of spurious topology changes in the Felsenstein zone, and thereby increases the accuracy of detecting recombination in DNA sequence alignments.

2.1 Methods

Consider an alignment $\mathcal{D}$ of m DNA sequences, N nucleotides long. Let each column in the alignment be represented by $\mathbf{y}_t$, where the subscript t represents the site, $1 \leq t \leq N$. Hence $\mathbf{y}_t$ is an m-dimensional column vector containing the nucleotides at the t site of the alignment, and $\mathcal{D} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$. Given a probabilistic model of nucleotide substitutions based on a homogeneous Markov chain with instantaneous rate matrix $\mathbf{Q}$, a phylogenetic tree topology S, and a vector of branch lengths $\mathbf{w}$, the probability of each column $\mathbf{y}_t$, $P(\mathbf{y}_t|S, \mathbf{w}, \theta)$, can be computed, as e.g. discussed in Husmeier *et al.* (2005a). Here, θ denotes a (vector) of free nucleotide substitution parameters extracted from $\mathbf{Q}$. For instance, for the

HKY85 model of Hasegawa *et al.* (1985), we have

$$\mathbf{Q} = \begin{pmatrix} . & \alpha\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha\pi_A & . & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & . & \alpha\pi_T \\ \beta\pi_A & \beta\pi_G & \alpha\pi_C & . \end{pmatrix} \tag{2.1}$$

where the dot in each row represents the additive inverse of the sum of the remaining elements in that row, $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$, with $\pi_i \in [0,1]$ and $\sum_i \pi_i = 1$, is a vector of nucleotide equilibrium frequencies, and $\alpha, \beta \geq 0$ are separate nucleotide substitution rates for transitions and transversions. For identifiability between $\mathbf{w}$ and $\mathbf{Q}$, the constraint $\sum_i Q_{ii}\pi_i = -1$ is commonly introduced defined in eq 1.29, which allows the branch lengths to be interpreted as expected numbers of mutations per site (see, e.g., Minin *et al.* (2005)). The normalisation constraint on π further reduces the number of free parameters by one, so that without loss of generality we have $\theta = (\pi_A, \pi_C, \pi_G, \tau)$, where $\tau = \alpha/\beta \geq 0$ is the transition-transversion ratio.

A Bayesian approach to phylogenetics without recombination was proposed and tested in Yang and Rannala (1997) and Larget and Simon (1999), where the objective is to sample the tree topology S, the branch lengths $\mathbf{w}$, and the parameters of the nucleotide substitution model, θ, from the posterior distribution $P(\mathbf{w}, S, \theta|\mathcal{D})$ with MCMC. Generalising this scheme to the presence of recombination requires replacing the single topology-indicating variable S by a sequence of topologies, $\mathbf{S} = (S_1, \ldots, S_N)$, where S_t (the 'state' at site t) represents the tree topology at site t. Each state $S_t \in \{1, \ldots, K\}$ can have a different vector of branch lengths, $\mathbf{w}_{S_t}$, and nucleotide substitution parameters, θ_{S_t}. To simplify the notation, we introduce the accumulated vectors $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_K)$ and $\theta = (\theta_1, \ldots, \theta_K)$ and define: $P(\mathbf{y}_t|S_t, \mathbf{w}_{S_t}, \theta_{S_t}) = P(\mathbf{y}_t|S_t, \mathbf{w}, \theta)$. This means that S_t indicates which subvectors of $\mathbf{w}$ and θ apply.

Since a tree topology may change as a result of recombination, which corresponds to a transition into another state S_t at the breakpoint t of the affected region, our main objective is the prediction of the state sequence $\mathbf{S} = (S_1, \ldots, S_N)$. This prediction should be based on the posterior probability $P(S_t|\mathcal{D})$, which requires a marginalisation over the other states

$$P(S_t|\mathcal{D}) = \sum_{S_1} \ldots \sum_{S_{t-1}} \sum_{S_{t+1}} \ldots \sum_{S_N} P(\mathbf{S}|\mathcal{D}) \tag{2.2}$$

and the remaining parameters to be integrating out:

$$P(\mathbf{S}|\mathcal{D}) = \int P(\mathbf{S}, \mathbf{w}, \theta|\mathcal{D}) d\mathbf{w} d\theta \tag{2.3}$$

Alternatively, if the objective is to detect only the location of recombination breakpoints without explicitly inferring the tree topologies in the different regions of the alignment, then the state sequences become nuisance parameters that have to be marginalised over. In practice this is effected by the introduction of a breakpoint detection operator, $\mathcal{B}$, which is a function of the state sequence, $\mathbf{S}$, and then obtaining the posterior probabilities of the breakpoints by summing over the state sequences:

$$P(\mathcal{B}|\mathcal{D}) \; = \sum_{\mathbf{S}} P(\mathcal{B}|\mathbf{S}) P(\mathbf{S}|\mathcal{D})$$

The assumption made for all three models discussed in Section 4 – MCP, DMCP and PFHMM – is that the integral over the branch lengths $\mathbf{w}$ can be solved analytically. We will revisit this point in Section 2.5, after briefly summarising the main ideas behind the three methods first.

2.2 Multiple change-point model (MCP)

In the MCP model, each state $S_t \in \{1, \ldots, K\}$ in $\mathbf{S} = (S_1, \ldots, S_N)$ represents a different tree topology. A separate vector of nucleotide substitution parameters $\theta_k, k \in \{1, \ldots, K\}$, and an overall divergence hyper parameter $\rho_k, k \in \{1, \ldots, K\}$, is associated with each state. As we will show later, in equation (2.9) and Section 2.5, the hyper parameter ρ_k defines the prior distribution of the branch lengths. The posterior probability is obtained from Bayes rule

$$P(\mathbf{S}, \theta, \rho, K|\mathcal{D}) \; \propto \; P(\mathcal{D}|\mathbf{S}, \theta, \rho, K) P(\mathbf{S}) P(\theta) P(\rho) P(K) \tag{2.4}$$

and requires the specification of various prior distributions. Note that the branch lengths $\mathbf{w}$ have been integrated out analytically. The prior on the number of states K is chosen to be a truncated Poisson distribution. For $P(\rho)$ a factorisable prior $P(\rho) = \prod_k P(\rho_k)$ is assumed, where each $P(\rho_k)$ is taken to be an exponential distribution. For $P(\theta)$ a similar factorisation is made: $P(\theta) = \prod_k P(\theta_k)$. The nucleotide substitution model chosen in Suchard *et al.* (2003) is the HKY85 model of (Hasegawa *et al.*, 1985), where the nucleotide equilibrium frequencies are kept

fixed, estimated from the whole DNA sequence alignment. Hence each θ_k corresponds to a single parameter, the transition-transversion ratio, and $P(\theta_k)$ is chosen to be the exponential distribution again. Finally, a change-point process is chosen as the prior on $P(\mathbf{S})$. The posterior probability over the state assignments is, in principle, obtained by marginalisation

$$P(\mathbf{S}|\mathcal{D}) = \sum_K \int\int P(\mathbf{S},\theta,\rho,K|\mathcal{D})d\theta d\rho \tag{2.5}$$

from which the prediction of topology changes is obtained by further marginalisation, e.g. according to equation (3.3). In practice, the integral in (2.5) is intractable and is approximated by sampling state sequences $\mathbf{S}$ and model parameters θ,ρ and K approximately from the posterior distribution of equation (2.4) with reversible jump Markov chain Monte Carlo (RJMCMC).

2.3 Dual multiple change-point model (DMCP)

A disadvantage of the MCP model is its inability to distinguish between recombination and rate heterogeneity. This shortcoming is addressed in the DMCP model of Minin *et al.* (2005), where two separate change point processes associated with the phylogenetic tree topology and the nucleotide substitution rate are employed. Let $\mathbf{S} = (S_1,\ldots,S_N)$ denote, as before, a hidden state sequence in which each state $S_t \in \{1,\ldots,K\}$ is associated with a phylogenetic tree topology. Denote by $\mathbf{R} = (R_1,\ldots,R_N)$ a separate hidden state sequence in which each hidden state $R_t \in \{1,\ldots,K'\}$ is associated with a divergence hyper parameter $\rho_{k'}, k' \in \{1,\ldots,K'\}$, and a (vector of) nucleotide substitution parameter(s) $\theta_{k'}, k' \in \{1,\ldots,K'\}$. Like $P(\mathbf{S})$, the prior on $\mathbf{R}$, $P(\mathbf{R})$, is chosen to be a change-point process, and both change-point processes are elected to be a priori independent: $P(\mathbf{S},\mathbf{R}) = P(\mathbf{S})P(\mathbf{R})$. The objective of Bayesian inference is to sample both hidden state sequences from the posterior distribution

$$P(\mathbf{S},\mathbf{R}|\mathcal{D}) = \sum_K \sum_{K'} \int\int P(\mathbf{S},\mathbf{R},K,K',\theta,\rho|\mathcal{D})d\theta d\rho \tag{2.6}$$

which is approximately effected with RJMCMC.

2.4 Phylogenetic factorial hidden Markov model (PFHMM)

The concept of the PFHMM of Husmeier (2005) is similar to the DMCP model. The main difference is the choice of the prior distribution on the hidden state sequences, $P(\mathbf{S},\mathbf{R}) = P(\mathbf{S})P(\mathbf{R})$. Rather than using two a priori independent change-point processes, two a priori independent homogeneous Markov chains are used. The details of the PFHMM can be found in section 1.9.4.

Note that the change-point process is a special case of a Markov chain, in which a state can only be visited once, without the possibility of a state reoccurring. This is an unnatural assumption in the context of recombination. When a recombination event has occurred in the central segment of a sequence alignment, then the evolutionary history of this central segment will be different from the flanking regions of the alignment. However, the two flanking regions share the same evolutionary history. This can be modelled with a Markov chain of two states and two transitions: from state 1 into state 2, and back from state 2 into state 1. However, a change-point process does not provide a mechanism to combine the two flanking regions into the same state. To rephrase this in terms of Markov chains: a change-point process corresponds to a Markov chain with two separate states for the two flanking regions, as the re-occurrence of a previously visited state is impossible. Consequently, the model has to infer the identity of the two states from the data. This is suboptimal, and it leads to an increased inference uncertainty (especially for short sequence alignments); see Lehrach (2008) for further details.

There are various differences in the detailed implementation of the methods. For the PFHMM described in Husmeier (2005), the parameters K, K', θ and ρ are fixed. This allows the computationally expensive RJMCMC simulations to be replaced by a much faster Gibbs sampling procedure. However, this difference is not essential to the PFHMM. In fact, the constraints on the parameters have been relaxed in Lehrach (2008) and Lehrach and Husmeier (2009), where – similarly to the work of Minin *et al.* (2005) – RJMCMC was used.

2.5 Analytic integration over the branch lengths

Consider a phylogenetic tree with topology S and branch lengths $\mathbf{w}$, denote the nucleotide substitution parameters by θ, and assume we are given a single column $\mathbf{y}$ from a DNA sequence alignment. The probability of this column, $\mathbf{y}$, is given by the following standard form (see, e.g., Husmeier *et al.* (2005a)):

$$P(\mathbf{y}|\mathbf{w},S,\theta) = \sum_{hidden} P(\tilde{y}_r)\prod_n P(\tilde{y}_n|\tilde{y}_{pa(n)},w^{pa(n)\to n},\theta) \tag{2.7}$$

Here, $\tilde{y}_n = y_n$ if the node n in the phylogenetic tree is observed (usually a leaf node). Otherwise, $\tilde{y}_n$ is a hidden variable (usually an ancestral node corresponding to a speciation point) that is marginalised over in the sum. The subscript r represents the root note, which for a reversible nucleotide substitution model can be chosen arbitrarily without affecting the probability of $\mathbf{y}$. The length of the branch connecting node n to its parent $pa(n)$ is denoted by $w^{pa(n)\to n}$. The factorisation in the expansion of equation (2.7) is defined by the phylogenetic tree topology S. We are interested in integrating out the branch lengths $\mathbf{w}$ according to

$$P(\mathbf{y}|S,\theta) = \int P(\mathbf{y}|\mathbf{w},S,\theta)P(\mathbf{w})d\mathbf{w} \tag{2.8}$$

We follow Suchard *et al.* (2003) and put a completely factorisable prior on the vector of branch lengths:

$$P(\mathbf{w}) = \prod_i P(w^i) = \frac{1}{\rho}\exp\left(-\frac{w^i}{\rho}\right) \tag{2.9}$$

where w^i is a single element of $\mathbf{w}$ representing the length of an individual branch connecting two nodes in the phylogenetic tree. Inserting this expression and equation (2.7) into equation (2.8) gives:

$$\begin{aligned} P(\mathbf{y}|S,\theta) &= \int \sum_{hidden} P(\tilde{y}_r)\prod_n P(\tilde{y}_n|\tilde{y}_{pa(n)},w^{pa(n)\to n},\theta)P(w^{pa(n)\to n})d\mathbf{w} \qquad (2.10)\\ &= \sum_{hidden} P(\tilde{y}_r)\prod_n \int P(\tilde{y}_n|\tilde{y}_{pa(n)},w^{pa(n)\to n},\theta)P(w^{pa(n)\to n})dw^{pa(n)\to n} \end{aligned}$$

Recall that $\tilde{y}_n$ and $\tilde{y}_{pa(n)}$ in $P(\tilde{y}_n|\tilde{y}_{pa(n)},w^{pa(n)\to n},\theta)$ represent nucleotides. The probability of nucleotide X mutating into Z along a branch of length w is of the following general form (Suchard *et al.*, 2003):

$$P(Z|X,w,\theta) = A_{ZX}\exp(-B_{ZX}w) + C_{ZX}\exp(-D_{XZ}w) + \pi_Z \tag{2.11}$$

Here, $A_{ZX}, B_{ZX}, C_{ZX}, D_{XZ}$ are nucleotide-dependent constants that are determined by the eigensystem of the instantaneous rate matrix $\mathbf{Q}$ and, thus, depend on the chosen nucleotide substitution model. For the HKY85 model, for instance, the particular expressions can be found in Hasegawa *et al.* (1985). The last term, π_Z, represents the equilibrium frequency of nucleotide Z, which is a parameter of the nucleotide substitution model. Hence, $A_{ZX}, B_{ZX}, C_{ZX}, D_{XZ}$ and π_Z are determined by θ.

Combining equations (2.9) and (2.11) allows the branch length to be integrated out analytically:

$$\begin{aligned} P(Z|X,\rho) &= \int P(Z|X,w)P(w|\rho)dw \\ &= \pi_Z + \frac{A_{ZX}}{\rho}\int \exp\left(-\left[B_{ZX}+\frac{1}{\rho}\right]w\right)dw \\ &\quad + \frac{C_{ZX}}{\rho}\int \exp\left(-\left[D_{XZ}+\frac{1}{\rho}\right]w\right)dw \\ &= \pi_Z + \frac{A_{ZX}}{1+B_{ZX}\rho} + \frac{C_{ZX}}{1+D_{XZ}\rho} \end{aligned} \tag{2.12}$$

Inserting eq. (2.12) into eq. (2.10) gives the following closed-form solution:

$$P(\mathbf{y}|S,\theta) = \sum_{hidden} P(\tilde{y}_r)\prod_n \left(\pi_{\tilde{y}_r} + \frac{A_{\tilde{y}_n,\tilde{y}_{pa(n)}}}{1+B_{\tilde{y}_n,\tilde{y}_{pa(n)}}\rho} + \frac{C_{\tilde{y}_n,\tilde{y}_{pa(n)}}}{1+D_{\tilde{y}_n,\tilde{y}_{pa(n)}}\rho}\right) \tag{2.13}$$

Let us now consider a whole DNA sequence alignment $\mathcal{D} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$:

$$P(\mathcal{D}|S,\theta) = \int P(\mathcal{D}|\mathbf{w},S,\theta)P(\mathbf{w})d\mathbf{w} = \int \prod_{t=1}^{N} P(\mathbf{y}_t|\mathbf{w},S,\theta)P(\mathbf{w})d\mathbf{w} \tag{2.14}$$

It is seen that the independence assumption of equation (2.9), $P(\mathbf{w}) = \prod_i P(w^i)$, does not yet allow this integral to be solved in closed form. What is needed is the expansion of the parameter space

$$\mathbf{w} \rightarrow (\mathbf{w}_1, \ldots, \mathbf{w}_t, \ldots, \mathbf{w}_N) \tag{2.15}$$

and the further independence assumption:

$$P(\mathbf{w}_1, \ldots, \mathbf{w}_t, \ldots, \mathbf{w}_N) = \prod_{t=1}^{N} P(\mathbf{w}_t) \tag{2.16}$$

Inserting this prior into eq. (2.14) gives

$$P(\mathcal{D}|S,\theta) = \prod_{t=1}^{N}\int P(\mathbf{y}_t|\mathbf{w}_t,S,\theta)P(\mathbf{w}_t)d\mathbf{w}_t = \prod_{t=1}^{N} P(\mathbf{y}_t|S,\theta) \tag{2.17}$$

where $P(\mathbf{y}_t|S,\boldsymbol{\theta})$ is given by (2.13). The commutation of the integral and the product, which is a direct consequence of equations (2.15) and (2.16), allows the integral to be solved in closed form, according to equations (2.10) and (2.13). The upshot is that for the branch lengths to be integrated out analytically, the model has to be modified so as to associate a separate branch length vector $\mathbf{w}_t$ with each position t in the DNA sequence alignment. This model is equivalent to the no-common-mechanism model proposed by Tuffley and Steel (1997). It is important to note that it is not the independence assumption of eq. (2.9) alone that leads to this simplification, a conclusion one might erroneously draw from Suchard *et al.* (2003). Rather, the more restrictive independence assumption of eq. (2.16) is needed. As a consequence of the latter independence assumption and the parameter expansion of eq. (2.15) the model is over-complex, though, with no information sharing between different sites with respect to the branch length estimation. In terms of statistical terminology, the expansion of eq. (2.15) turns the structural parameters $\mathbf{w}$ into a set of incidental parameters[1]. As discussed in Goldman (1990), this implies that maximum likelihood is no longer guaranteed to provide a consistent estimator. This aspect, which has not been considered in any of the three methods discussed in Section 4 – MCP, DMCP and PFHMM – causes inconsistency problems that are related to those found in maximum parsimony. We will investigate them more closely in the subsequent sections.

2.6 Data

We suspect that the assumption of independent site-specific branch lengths, as discussed in Section 2.5, could lead to inconsistency problems akin to those that affect maximum parsimony (Felsenstein, 1978b). To test this conjecture, we tested the models on synthetic DNA sequence alignments. We used two different programs for generating these alignments: SEQGEN (Rambaut, 1996) and the MATLAB programs used in Husmeier (2005). In both cases we simulated the nucleotide substitution processes with the HKY model (Hasegawa *et al.*, 1985), using a transition-transversion ratio of 2 and uniform nucleotide equilibrium frequencies. For SEQGEN, we used the implementation available via the web service

[1] Structural parameters are parameters that appear in the probability distributions of all the observations, whereas an incidental parameter appears in the probability distributions of only a subset of the observations. See Goldman (1990) for further details.

provided by the Pasteur Institute, available from

$$http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?form=seqgen.$$

The MATLAB programs used in Husmeier (2005) are available from

$$http://www.bioss.ac.uk/staff/dirk/Supplements/,$$

and were preferred when running a large number of jobs in batch mode. We generated two different types of alignments: homogeneous alignments, and alignments with mosaic structures.

2.6.1 Homogeneous DNA sequence alignment

A homogeneous DNA sequence alignment is an alignment where one single phylogenetic tree with a specified branch length vector is used in the data generating process. We generated alignments from the 4-taxa tree depicted in Figure 2.1 for different settings of the branch length configurations, specified by the values $d2$ and $d3$. This corresponds to a study originally carried out by Felsenstein for investigating potential shortcomings and inconsistencies of maximum parsimony (Felsenstein, 1978b). We varied the parameters $d2$ and $d3$, defined in Figure 2.1, over a large range that included the so-called Felsenstein zone, in which maximum parsimony systematically fails. The data thus generated were used for the studies reported in Figures 2.4, 2.5, 2.6, and 2.8. All sequence alignments were 1000 nucleotides long.

2.6.2 DNA sequence alignment with mosaic structure

A mosaic structure is a DNA sequence alignment subject to recombination and/or rate heterogeneity, where a segment in the DNA sequence alignment was generated from a tree with a different tree topology, or with different branch lengths. We generated DNA sequence alignments from the 4-taxa tree shown in Figure 2.1. The alignments were 1500 nucleotides long. They contained a central segment of 500 nucleotides, which was generated from a tree with the same topology as for the flanking regions, but with a different branch length configuration. The objective of our study was to investigate if spurious tree topology changes were inferred with the recombination detection methods described in Section 2.1 if the branch length configurations for the central and flanking regions were on different

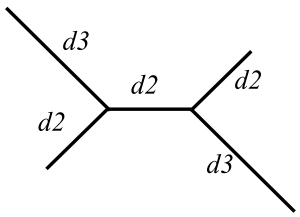


Figure 2.1: Phylogenetic tree of four taxa

The figure shows a phylogenetic tree of four taxa, which was used for generating the synthetic DNA sequence alignments, as described in Section 3.3. The tree contains two types of branch lengths, denoted by $d2$ and $d3$, as in Felsenstein (1978b). For configurations with large branch lengths $d3$ and small branch lengths $d2$, the method of maximum parsimony is known to systematically infer the wrong tree topology.

sides of the Felsenstein boundary. The alignments thus generated were used in the study described in the caption of Figure 2.9.

2.7 Bayesian model selection

As discussed in Section 2.5, the integration over the branch lengths, on which the three methods MCP, DCMP and PFHMM rely, is based on the choice of independent site-specific branch lengths, that is, the vector of branch lengths is allowed to be different at each site. Since separate branch length vectors $\mathbf{w}_t$ are associated with different positions t in the alignment, there is no longer a mechanism in place to over-rule a posteriori the prior independence assumption of equation (2.9). We suspect that MCP, DCMP and PFHMM might therefore be susceptible to the same inconsistency problems as the method of maximum parsimony, which could result in the prediction of spurious topology changes. Before investigating this conjecture in direct simulation studies, to be discussed in Section 2.12.2, we carried out systematic Bayesian model selection along the Felsenstein zone (introduced in Felsenstein (1978b)). To this end, we generated data synthetically from the four-taxa tree of Figure 2.1 with two types of branches, $d2$ and $d3$, as described in Section 3.3.

For the different $d2/d3$ ratios, we estimated the marginal likelihood $P(\mathcal{D}|S,\mathcal{H}_i)$ for each of the three possible tree topologies, $S \in \{\Psi_1, \Psi_2, \Psi_3\}$, under the two hypotheses or modelling approaches: independent site-specific branch length vectors

$\mathbf{w}_t$ ($\mathcal{H}_0$), and a common vector of branch lengths $\mathbf{w}$ for the whole alignment ($\mathcal{H}_1$). Under the assumption of a uniform prior on the tree topologies, we estimated the posterior probability for the correct tree topology

$$P(S = true|\mathcal{D}, \mathcal{H}_i) = \frac{P(\mathcal{D}|S = true, \mathcal{H}_i)}{\sum_{k=1}^{3} P(\mathcal{D}|S = \Psi_k, \mathcal{H}_i)} \tag{2.18}$$

We investigated the behaviour of $P(S = true|\mathcal{D}, \mathcal{H}_i)$ in $d2/d3$ space, especially around the Felsenstein zone. For estimating the marginal likelihood, we pursued two approaches: an inter-model approach, using MCMC, and an intra-model approach, using the method of annealed importance sampling (AIS), as proposed in Neal (2001). Below, in Sections 2.8 and 2.9, we will first define the exact form of the probabilistic models associated with $\mathcal{H}_0$ and $\mathcal{H}_1$. We will then, in Sections 2.10 and 2.11, briefly describe the way we computed the marginal likelihoods. Finally, we will present the results and investigate the behaviour of the two modelling frameworks around the Felsenstein zone.

2.8 Independent site-specific branch-length model $\mathcal{H}_0$

To investigate whether there are potential inconsistency problems inherent in the recombination detection methods MCP, DMCP and PFHMM, we considered the standard phylogenetic model (A reminder that the standard model assumes a single topology along the sequence alignment without recombination) subject to the same independence assumptions of equations (2.9) and (2.16) on which MCP, DMCP and PFHMM are based. We refer to this modelling concept as $\mathcal{H}_0$. The corresponding graphical model is shown in Figure 2.2. The right panel depicts the site-independence of the branch lengths, inherent in equation (2.16). As a consequence of the analytic integration over the branch lengths, discussed in Section 2.5, the model simplifies. The resulting probabilistic graphical model is shown in Figure 2.2a, which defines the following factorisation:

$$P(\mathcal{D}, S, \rho, \alpha) = P(\mathcal{D}|S, \rho)P(S)P(\rho|\alpha) \tag{2.19}$$

Here, $\mathcal{D}$ is the DNA sequence alignment, S is the tree topology, ρ (defined in equation (2.9)) represents the average mutational divergence, and α is a hyper parameter that defines the prior distribution of ρ:

$$P(\rho|\alpha) = \frac{1}{\alpha} e^{-\frac{\rho}{\alpha}} \tag{2.20}$$

The prior distribution over tree topologies, $P(S)$, is chosen to be uniform. The objective of Bayesian model selection for learning the best tree topology S is to estimate the marginal likelihood

$$P(\mathcal{D}|S,\alpha) = \int P(\mathcal{D}|S,\rho)P(\rho|\alpha)d\rho, \tag{2.21}$$

which is the numerator in the model selection equation (2.18). The term $P(\mathcal{D}|S,\rho)$ is given in (2.17), where the explicit reference to $\mathcal{H}_0$ and the nucleotide substitution parameters θ has been left out to reduce the notational complexity.[2]

2.9 Standard phylogenetic model $\mathcal{H}_1$

For comparison with $\mathcal{H}_0$, we consider the standard phylogenetic model, in which a common vector of branch lengths $\mathbf{w}$ is used for the whole DNA sequence alignment, as depicted in Figure 2.3b. We refer to this modelling concept as $\mathcal{H}_1$. The essential difference from $\mathcal{H}_0$ is that the independence assumption of equation (2.16), on which MCP, DMCP and PFHMM are based, is no longer valid. The consequence is that the elimination of the branch lengths, as described in Section 2.5 and represented in Figure 2.2a, is no longer feasible, resulting in the more complex probabilistic dependence model of Figure 2.3a. The structure of the model incorporates the average mutational divergence ρ, and the branch length vector $\mathbf{w}$, and the joint probability factorizes as follows:

$$P(\mathcal{D},S,\mathbf{w},\rho,\alpha) = P(\mathcal{D}|S,\mathbf{w})P(\mathbf{w}|\rho)P(S)P(\rho|\alpha) \tag{2.22}$$

$P(S)$ is the prior distribution over tree topologies, which we keep uniform. The prior distribution over branch lengths, $P(\mathbf{w}|\rho)$, is defined in equation (2.9). This distribution depends on the hyper parameter ρ, which is given the prior distribution of equation (2.20). The objective of Bayesian model selection is to estimate the marginal likelihood

$$P(\mathcal{D}|S,\alpha) = \int P(\mathcal{D}|S,\mathbf{w})P(\mathbf{w}|\rho)P(\rho|\alpha)d\rho d\mathbf{w} \tag{2.23}$$

[2]Recall that the free nucleotide substitution parameters θ of the HKY model are the equilibrium frequencies and the transition-transversion ratio. In our simulations, we chose a uniform distribution for the equilibrium frequencies, and a fixed transition-transversion ratio of 2. Also, note that in (2.17) the explicit reference to the hyperparameter ρ has been dropped for notational convenience.

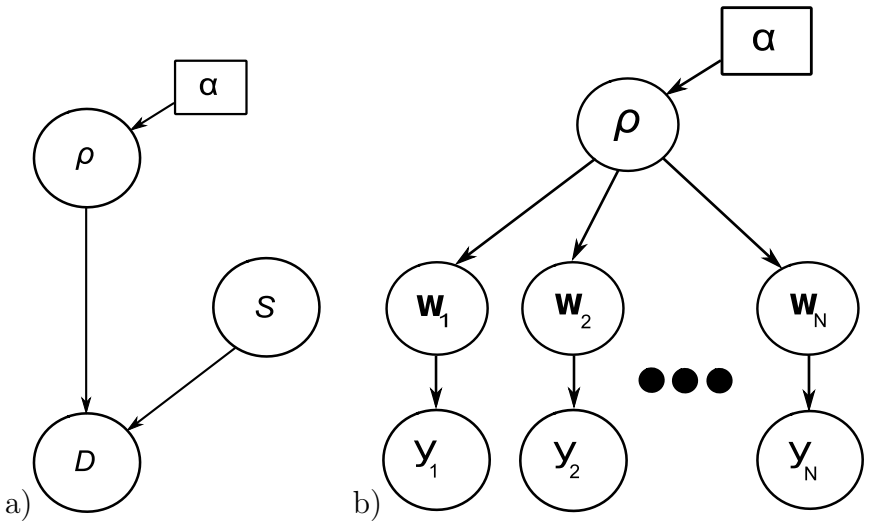


Figure 2.2: Graphical model for hypothesis $\mathcal{H}_0$

Hypothesis $\mathcal{H}_0$ is based on the independence assumption of equation (2.16), which is depicted in Panel b). Here, the $\mathbf{y}_t$ represent the columns in the DNA sequence alignment, the $\mathbf{w}_t$'s are separate independent vectors of branch lengths, associated with the sites t in the alignment, and ρ is a hyper parameter determining the prior distribution over the branch lengths, via equation (2.9). As a consequence of the independence assumptions inherent in this model, the branch lengths can be integrated out, as described in Section 2.5. The resulting probabilistic graphical model is shown in Panel a). Here D (which is equal to $\mathcal{D}$ in the text) is the DNA sequence alignment, S is the tree topology, ρ (defined in equation (2.9)) represents the average mutational divergence, and α is a hyper parameter that defines the prior distribution of ρ; see equation (2.20). Further details are given in Section 2.8.

Here, $P(\mathcal{D}|S,\mathbf{w})$ is the (non-marginal) likelihood, which is obtained from $P(\mathbf{y}_t|S,\mathbf{w})$ defined in equation (2.7) as follows:

$$P(\mathcal{D}|S,\mathbf{w}) = \prod_{t=1}^{N} P(\mathbf{y}_t|S,\mathbf{w}) \tag{2.24}$$

Note that in order to simplify the notation and the graphical presentation, we have not made the dependence on the nucleotide substitution parameters θ explicit in equation (2.24) and Figures 2.2-2.3.

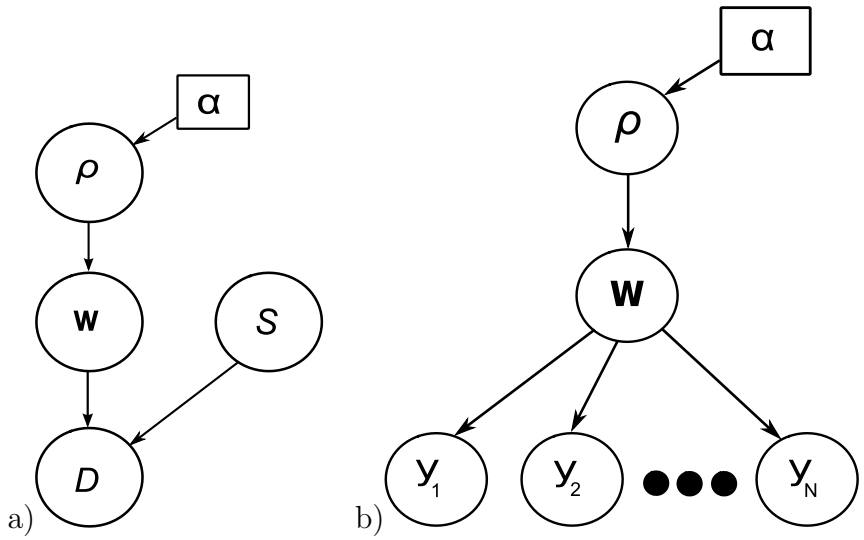


Figure 2.3: Graphical model for hypothesis $\mathcal{H}_1$

The symbols are the same as those defined in Figure 2.2. Panel b) shows that a common vector of branch lengths **w** is used to describe the whole DNA sequence alignment, rather than independent site-specific vectors, as in Figure 2.2b. The consequence is that the elimination of the branch lengths, as described in Section 2.5 and represented in Figure 2.2a, is no longer feasible, resulting in the more complex probabilistic dependence model of Panel a). Further details are given in Section 2.9.

2.10 Inter-model approach: Markov chain Monte Carlo (MCMC)

The objective of the inter-model approach is to sample tree topologies from the posterior distribution of equation (2.18).

2.10.1 MCMC framework for hypothesis $\mathcal{H}_0$

Recall that for a DNA sequence alignment with four sequences, there are three different unrooted tree topologies. Our proposal distribution for proposing a new tree topology S^* from the current topology S is just the uniform distribution over the tree topology space. For proposing a new rate[3] ρ^*, we sample a value $\rho^\sharp$ from the uniform interval of length W centred on the current value ρ, using reflection to ensure the proposed value is positive: $\rho^* = \rho^\sharp$ if $\rho^\sharp \geq 0$; otherwise $\rho^* = -\rho^\sharp$.

[3]In a slight abuse of terminology, we henceforth refer to the hyper parameter ρ as the "rate".

This proposal distribution depends on a tuning parameter W, which is adjusted during the burn-in period to achieve a target acceptance rate between 30% and 70%. The Metropolis-Hastings acceptance probability for this move is:

$$a(S^*,\rho^*|S,\rho) = \min\left\{1, \frac{Q(\rho|\rho^*)P(\rho^*|\alpha)Q(S|S^*)P(S^*)P(\mathcal{D}|S^*,\rho^*)}{Q(\rho^*|\rho)P(\rho|\alpha)Q(S^*|S)P(S)P(\mathcal{D}|S,\rho)}\right\} \tag{2.25}$$

where $P(\mathcal{D}|S,\rho)$ is the likelihood, defined in equation (2.17), $P(S)$ and $P(\rho|\alpha)$ are the prior distributions for the tree topology and the rate, defined in Section 2.8, and $Q(S^*|S)$ and $Q(\rho^*|\rho)$ are the proposal distributions, as discussed above. It is straightforward to show that the latter distributions are symmetric and thus cancel out. The prior distribution in tree topology space, $P(S)$, is uniform. Equation (2.25) thus simplifies as follows:

$$a(S^*,\rho^*|S,\rho) = \min\left\{1, \frac{P(\rho^*|\alpha)P(\mathcal{D}|S^*,\rho^*)}{P(\rho|\alpha)P(\mathcal{D}|S,\rho)}\right\} \tag{2.26}$$

To increase the acceptance probabilities, we de-couple the proposal step into two separate steps for proposing a new tree topology and a new rate, with the following acceptance probabilities:

$$a(S^*|S) = \min\left\{1, \frac{P(\mathcal{D}|S^*,\rho)}{P(\mathcal{D}|S,\rho)}\right\}, \tag{2.27}$$

$$a(\rho^*|\rho) = \min\left\{1, \frac{P(\rho^*|\alpha)P(\mathcal{D}|S,\rho^*)}{P(\rho|\alpha)P(\mathcal{D}|S,\rho)}\right\}. \tag{2.28}$$

2.10.2 MCMC framework for hypothesis $\mathcal{H}_1$

Recall that for Hypothesis $\mathcal{H}_1$ the analytic integration over the branch lengths $\mathbf{w}$ is no longer tractable; hence, the sampling of a new vector of branch lengths $\mathbf{w}^*$ from the existing branch lengths $\mathbf{w}$ has to be incorporated into the MCMC scheme. We elected to propose new values $w_i^\sharp$ independently from a Cauchy distribution centred on the current value w_i

$$Q(w_i^\sharp|w_i,\gamma) = \frac{1}{\pi\gamma\left(1+\left(\frac{w_i^\sharp - w_i}{\gamma}\right)^2\right)} \tag{2.29}$$

subject to the constraint that the proposed new branch length w_i^* must be non-negative. Again, this constraint is achieved by reflection: $w_i^* = w_i^\sharp$ if $w_i^\sharp \geq 0$; otherwise $w_i^* = -w_i^\sharp$.

The technique used to draw samples from the Cauchy distribution is by generating random numbers on the interval $[0,1]$ and mapping these to the values on the CDF of the Cauchy,

$$\frac{1}{\pi}\arctan\left(\frac{w_i^\sharp - w_i}{\gamma}\right)+\frac{1}{2}. \tag{2.30}$$

The spread of the proposal distribution is defined by the tuning parameter γ, which is adjusted during the burn-in phase to achieve an average acceptance rate between 30% and 70%. The Cauchy distribution is chosen for the thick tails that it has. To avoid local optima a series of sequential samples can be used to traverse the space, but samples very distant from the present point can also be generated occasionally to propose alternative high likelihood points with this distribution. These occasional distant points being proposed assisted convergence to the true posterior distribution.

The proposal distributions for the tree topology S and the rate ρ are the same as discussed in the previous subsection. The Metropolis-Hastings acceptance probability is given by

$$a(S^*,\rho^*,\mathbf{w}^*|S,\rho,\mathbf{w}) = \min\{1,r\} \tag{2.31}$$

$$r = \frac{Q(\rho|\rho^*)P(\rho^*|\alpha)\left(\prod_{i=1}^K Q(w_i|w_i^*)P(w_i^*|\rho)\right)Q(S|S^*)P(S^*)P(\mathcal{D}|S^*,\mathbf{w}^*)}{Q(\rho^*|\rho)P(\rho|\alpha)\left(\prod_{i=1}^K Q(w_i^*|w_i)P(w_i|\rho)\right)Q(S^*|S)P(S)P(\mathcal{D}|S,\mathbf{w})}$$

where $K=\dim\{\mathbf{w}\}$, $Q(w_i^*|w_i)$ is the proposal distribution for a new branch length, which is straightforward to compute from equation (2.29) and the condition of reflection, $P(w_i^*|\rho)$ is the prior distribution of the branch lengths, defined in equation (2.9). $P(\mathcal{D}|S,\mathbf{w})$ is defined in equation (2.24). The other expressions are the same as defined below equation (2.25) in the previous subsection.

It is straightforward to show that the proposal distribution for the branch lengths, $Q(w_i^*|w_i)$, is symmetric and thus cancels out. Together with the simplifications discussed below equation (2.25) we get the following simplified expression:

$$a(S^*,\rho^*|S,\rho) = \min\left\{1,\frac{P(\rho^*|\alpha)\prod_{i=1}^K P(w_i^*|\rho)P(\mathcal{D}|S^*,\mathbf{w}^*)}{P(\rho|\alpha)\prod_{i=1}^K P(w_i|\rho)P(\mathcal{D}|S,\mathbf{w})}\right\} \tag{2.32}$$

As with the model discussed in the previous subsection, we de-couple the individual update steps so as to increase the acceptance probability:

$$a(\mathbf{w}^*|\mathbf{w}) = \min\left\{1,\frac{\prod_{i=1}^K P(w_i^*|\rho)P(\mathcal{D}|S,w_i^*)}{\prod_{i=1}^K P(w_i|\rho)P(\mathcal{D}|S,w_i)}\right\} \tag{2.33}$$

$$a(\rho^*|\rho) = \min\left\{1, \frac{P(\rho^*|\alpha)\prod_{i=1}^{K} P(w_i|\rho^*)}{P(\rho|\alpha)\prod_{i=1}^{K} P(w_i|\rho)}\right\} \tag{2.34}$$

$$a(S^*|S) = \min\left\{1, \frac{P(\mathcal{D}|\mathbf{w}, S^*)}{P(\mathcal{D}|\mathbf{w}, S)}\right\} \tag{2.35}$$

2.10.3 Convergence diagnostics

The Gelman and Rubin diagnostic test (Gelman and Rubin, 1992) was used in the simulations for both models to investigate whether the chains have converged. The tests output a (set of) so-called potential scale reduction factor(s) (PSRF), where a value close to 1 provides a strong indication of convergence. We computed the PSRF from the branch lengths and the rate hyper parameter ρ, and chose burn-in and simulation lengths that led to PSRFs below 1.1. This was effected with the following settings. For $\mathcal{H}_0$, we carried out 2K burn-in and 5K sampling steps. For $\mathcal{H}_1$, these values had to be slightly increased (owing to the larger dimension of the parameter space), to 5K burn-in and 10K sampling steps.

2.11 Intra-model approach: Annealed importance sampling (AIS)

As an alternative to the inter-model MCMC sampling scheme discussed in the previous section, we consider an intra-model approach, where the objective is a direct (approximate) computation of the marginal likelihood

$$P(\mathcal{D}|S) = \int P(\mathcal{D}|\phi, S)P(\phi|S)d\phi \tag{2.36}$$

where ϕ is the vector of all parameters associated with the respective hypothesis: $\phi = \rho$ under the site-independent branch length hypothesis $\mathcal{H}_0$, and $\phi = (\mathbf{w}, \rho)$ for $\mathcal{H}_1$. In principle one could approximate the marginal likelihood by

$$P(\mathcal{D}|S) \approx \frac{1}{N}\sum_{t=1}^{N} P(\mathcal{D}|\phi_t, S) \tag{2.37}$$

where $\{\phi_t\}$ is a sample from the prior distribution $P(\phi|S)$. However, the convergence of this estimator is known to be poor unless the prior and posterior distributions are very similar (Raftery, 1996). Alternatively, one could exploit the

Bayesian identity $P(\mathcal{D}|\phi,S)P(\phi|S) = P(\phi|\mathcal{D},S)P(\mathcal{D}|S)$ and compute the marginal likelihood from the so-called harmonic mean estimator (Raftery, 1996)

$$\frac{1}{P(\mathcal{D}|S)} \approx \frac{1}{N}\sum_{t=1}^{N}\frac{1}{P(\mathcal{D}|\phi_t,S)} \tag{2.38}$$

using a sample $\{\phi_t\}$ from the posterior distribution $P(\phi|\mathcal{D},S)$. This estimator is known to be numerically unstable, since for modestly informative priors the main contributions to the sum on the right-hand side of equation (2.38) come from the tail rather than the bulk of the posterior distribution. The standard approach to deal with these problems is to use importance sampling. Define some (possibly unnormalized) distribution $Q(\phi)$, and rewrite equation (2.36) in the form:

$$\frac{P(\mathcal{D}|S)}{Z_Q} = \int \frac{P(\mathcal{D}|\phi,S)P(\phi|S)}{Q(\phi)}\frac{Q(\phi)}{Z_Q}d\phi \tag{2.39}$$

where $Z_Q = \int Q(\phi)d\phi$. Provided $Q(\phi) \neq 0$ whenever $P(\mathcal{D}|\phi,S)P(\phi|S) \neq 0$, we get the following unbiased and consistent estimator of the marginal likelihood (Neal, 2001):

$$\frac{P(\mathcal{D}|S)}{Z_Q} \longleftarrow \frac{1}{N}\sum_{t=1}^{N} c_t \tag{2.40}$$

where $\{\phi_t\}$ is a sample drawn from $\frac{Q(\phi)}{Z_Q}$, and the weights c_t are defined as $c_t = \frac{P(\mathcal{D}|\phi_t,S)P(\phi_t|S)}{Q(\phi_t)}$. Rather than using some fixed distribution $Q(\phi)$ as a compromise between the prior and the posterior distribution, as in Raftery (1996), we follow the annealed importance sampling (AIS) scheme proposed in Neal (2001), where the idea is to propose new values $\{\phi_t\}$ by gradually transforming the prior into the posterior distribution. Define

$$Q_m(\phi) = P(\phi|S)^{[1-\beta_m]}P(\phi|\mathcal{D},S)^{\beta_m} \tag{2.41}$$

where $1 = \beta_0 > \beta_1 > \ldots > \beta_M = 0$. That is, Q_M is equal to the prior, and Q_0 is equal to the posterior distribution. AIS produces a sample of parameter vectors $\{\phi_t\}$ and associated weights $\{c_t\}$ according to the following procedure. Consider a Markov chain transition defined by $T_m(\mathbf{x}'|\mathbf{x})$ giving the probability of moving from the current state $\mathbf{x}$ to the new state $\mathbf{x}'$. The choice of T_m is decided by the requirement that it must leave the corresponding probability distribution Q_m in equation (2.41) invariant, e.g. by satisfying the equation of detailed balance: $T_m(\mathbf{x}'|\mathbf{x})Q_m(\mathbf{x}) = T_m(\mathbf{x}|\mathbf{x}')Q_m(\mathbf{x}')$. Next, a sequence of points is generated as fol-

lows:

$$\begin{aligned}
&\text{Generate } \mathbf{x}_{M-1} \text{ from } Q_M \\
&\text{Generate } \mathbf{x}_{M-2} \text{ from } \mathbf{x}_{M-1} \text{ using } T_{M-1} \\
&\dots \\
&\text{Generate } \mathbf{x}_1 \text{ from } \mathbf{x}_2 \text{ using } T_2 \\
&\text{Generate } \mathbf{x}_0 \text{ from } \mathbf{x}_1 \text{ using } T_1
\end{aligned} \tag{2.42}$$

The proposed parameter vector of the tth iteration is set to $\phi_t = \mathbf{x}_0$, and the associated weight is set to

$$c_t = \frac{Q_{M-1}(\mathbf{x}_{M-1})}{Q_M(\mathbf{x}_{M-1})} \frac{Q_{M-2}(\mathbf{x}_{M-2})}{Q_{M-1}(\mathbf{x}_{M-2})} \dots \frac{Q_1(\mathbf{x}_1)}{Q_2(\mathbf{x}_1)} \frac{Q_0(\mathbf{x}_0)}{Q_1(\mathbf{x}_0)} \tag{2.43}$$

The scheme is continued to generate a sample of weights $\{c_t\}$. It can be shown that for the sample of weights thus obtained, equation (2.40) provides a consistent and unbiased estimator of the marginal likelihood (Neal, 2001). The individual steps of (2.42) can be constructed by applying the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953) to the respective transition probability T_m, as in MCMC. Note that as opposed to MCMC, the respective Markov chains do not need to be run to convergence, though.

In our simulations, we carried out for each step in (2.42) 10 Metropolis-Hastings steps according to the description in Section 2.10.1, for $\mathcal{H}_0$, and Section 2.10.2, for $\mathcal{H}_1$. A "temperature" ladder of $M = 10$ equidistant β_m values for defining the intermediate distributions Q_m in (2.41) was selected, and we chose a total sample size of $N = 400$ for computing the marginal likelihood according to (2.37). We experimented with a polynomial rather than an equidistant cooling scheme for β, but did not find any noticeable differences in the results.

As a heuristic indicator of how accurate the estimation with AIS is, we followed Neal (2001) and computed the variance of $c_t^* = c_t / \frac{1}{N} \sum_{t=1}^{N} c_t$. The term $\psi = 1/[1 + Var(c_t^*)]$ gives a rough indication of the factor by which the sample size is effectively reduced when drawing samples according to the procedure (2.42) rather than from the correct posterior distribution. In our simulations, we typically found values of $\psi \leq 1.3$, indicating a sufficient degree of convergence.

2.12 Results

2.12.1 Investigating the behaviour around the Felsenstein zone

We generated synthetic DNA sequence alignments from 4-taxa trees with different branch lengths. In the vein of Felsenstein's seminal study for demonstrating the inconsistency of maximum parsimony (Felsenstein, 1978b), we systematically varied the parameters $d2$ and $d3$ in Figure 2.1, and generated DNA sequence alignments with SEQGEN (Rambaut, 1996), as described in Section 3.3. For each branch length configuration $[d2, d3]$, we estimated the posterior probabilities for the three possible tree topologies under the two different models discussed above: the site-specific branch length model $\mathcal{H}_0$, described in Section 2.8, and the standard phylogenetic model $\mathcal{H}_1$, described in Section 2.9. We repeated the estimation of the posterior probabilities with two different methods: MCMC, as described in Section 2.10, and annealed importance sampling, as described in Section 2.11. The results are shown in Figures 2.4 and 2.5. In both figures, subfigure a) shows the results for the site-specific branch length model $\mathcal{H}_0$, while subfigure b) shows the results for the standard phylogenetic model $\mathcal{H}_1$. The axes represent the values of the parameters $d2$ and $d3$; hence each grid location defines a phylogenetic tree with a specific branch length configuration. The estimated posterior probabilities are indicated with a grey shading ranging from 0 (black) to 1 (white) and the values in between are indicated by the legend in subfigure c). It is clearly seen that the independent branch length model $\mathcal{H}_0$ leads to a systematic failure in the Felsenstein zone (characterized by a small value of $d2$ and a large value of $d3$) in that the posterior probability of the correct tree is consistently close to zero. In fact, the tree topology with the highest posterior probability was found to be the one in which the two longer branches were grouped together. This suggests that the independent branch length model $\mathcal{H}_0$ has the same problem with long-branch attraction as maximum parsimony. This failure was avoided with the standard phylogenetic model $\mathcal{H}_1$, whose posterior probability of the correct tree topology was consistently above 0.5 (and mostly close to 1) for the whole branch length configurations space. The results obtained with MCMC and AIS were largely consistent, although the difference between the posterior probabilities for $\mathcal{H}_0$ and $\mathcal{H}_1$ in the Felsenstein zone was slightly larger with MCMC than with

AIS. In terms of performance between the inter and intra model approaches, simulations with AIS required less CPU time for convergence but more frequently missed the posterior distribution. From carefully monitoring simulations done by both methods it appears that AIS was more sensitive to the samples proposed by the proposal distribution than MCMC. The short MCMC simulations at the intermediate temperatures of the annealing procedure did not explore the space as much as the inter model approach did especially at low temperatures. This assisted in convergence once a posterior was found, but in the case where this was not the desired poster it was rare for it to change modes.

2.12.2 Evaluation of the performance of DMCP and PFHMM

The previous section has shown that for the model of independent, site-specific branch lengths ($\mathcal{H}_0$), there is a systematic failure in the Felsenstein zone, which is avoided with the standard phylogenetic model of common branch lengths, $\mathcal{H}_1$. Since the recombination detection methods PFHMM and DMCP are based on $\mathcal{H}_0$, we suspect that they are susceptible to the same systematic failure. We tested this conjecture by applying both methods, DMCP and PFHMM, to the same synthetic DNA sequence alignments as used in the previous section. For comparison, we also applied the phylogenetic hidden Markov model (PHMM) of Husmeier and McGuire (2003). Note that the latter model is based on $\mathcal{H}_1$ and should therefore not be susceptible to inferring wrong tree topologies in the Felsenstein zone.[4] We used the authors' own programs, available from the webpages referenced in Minin *et al.* (2005) (for DMCP), Husmeier (2005) (for PFHMM) and Husmeier and McGuire (2003) (for PHMM). All three methods sample parameters and hidden states from the posterior distribution with MCMC. To test for convergence of these simulations, we computed the potential scale reduction factor from different quantities, as in Gelman and Rubin (1992), taking values below 1.1 as an indication of sufficient convergence[5]. From the sampling phase of the MCMC simulations, we computed for each site t in the alignment

[4]Note that as opposed to DMCP and PFHMM, PHMM cannot distinguish between recombination and rate heterogeneity, though.

[5]This was achieved with the following burn-in and sampling lengths. Burn-in: 1000 steps for DMPC, 250 steps for PFHMM, and 10K steps for PHMM. Sampling phase: 200 subsample steps (in intervals of 50 steps) for DMPC, 250 steps for PFHMM, and 1000 subsample steps (in intervals of 10 steps) for PHMM. PFHMM needs as input a set of fixed nucleotide substitution rates, corresponding to the hyperparameter ρ in eq 2.9. These values were selected as $\rho \in \{0.05, 0.1, 0.5, 1, 2, 4, 6, 8\}$.

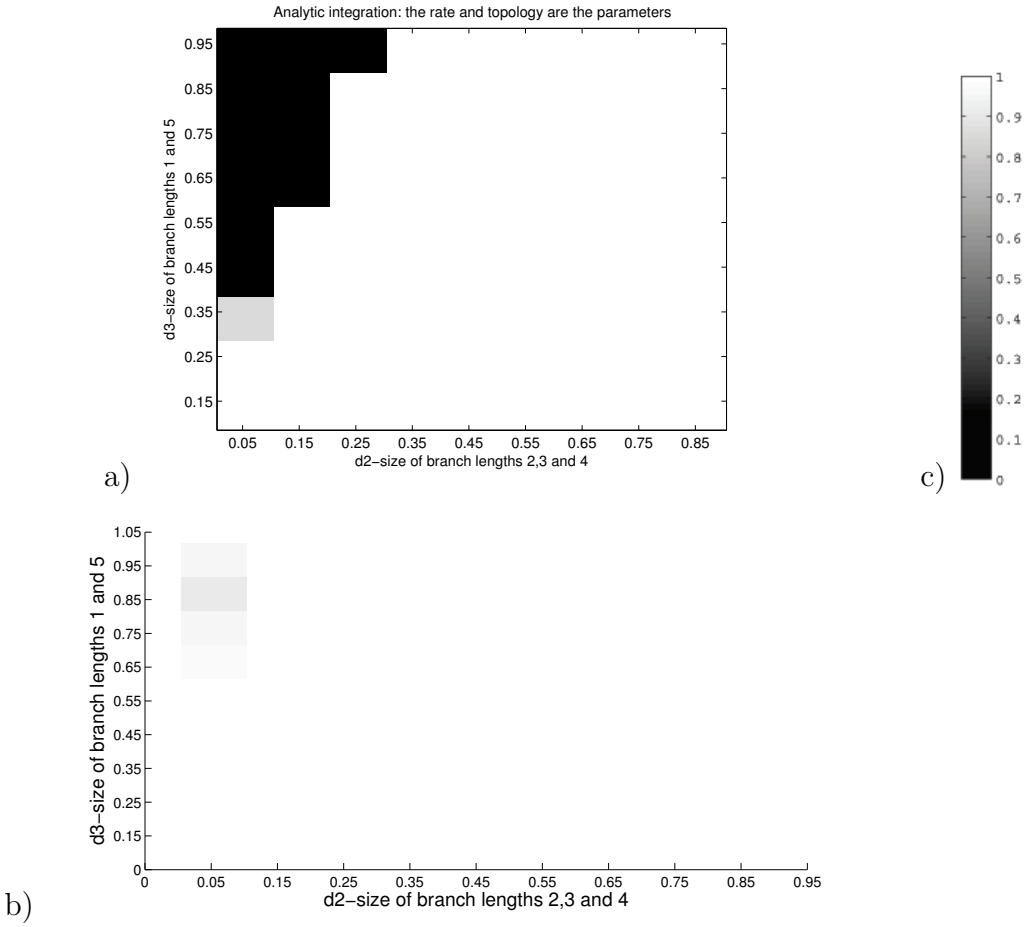


Figure 2.4: Posterior probabilities estimated with MCMC.

The two figures show the posterior probability of the correct tree topology for different branch length configurations. These configurations are determined by the values of $d2$ and $d3$, as defined in Figure 2.1. In each subfigure, the horizontal axis refers to $d2$, and the vertical axis refers to $d3$. The grey shading indicates the value of the inferred posterior probability, as indicated in the legend on the right, ranging from 0 (black) to 1 (white). Subfigure a) shows the results obtained for Model $\mathcal{H}_0$, represented in Figure 2.2. Subfigure b) shows the results obtained for Model $\mathcal{H}_1$, represented in Figure 2.3. The results shown are those obtained from a specific set of DNA sequence alignments generated from trees with the indicated $[d2, d3]$ configurations, as described in Section 2.6.1. Repeating the simulations for different sequences generated from the same trees was found to give nearly identical results. It is clearly observed that Model $\mathcal{H}_0$, which is shown in Subfigure a), leads to the systematic prediction of the wrong tree topology in the Felsenstein zone.

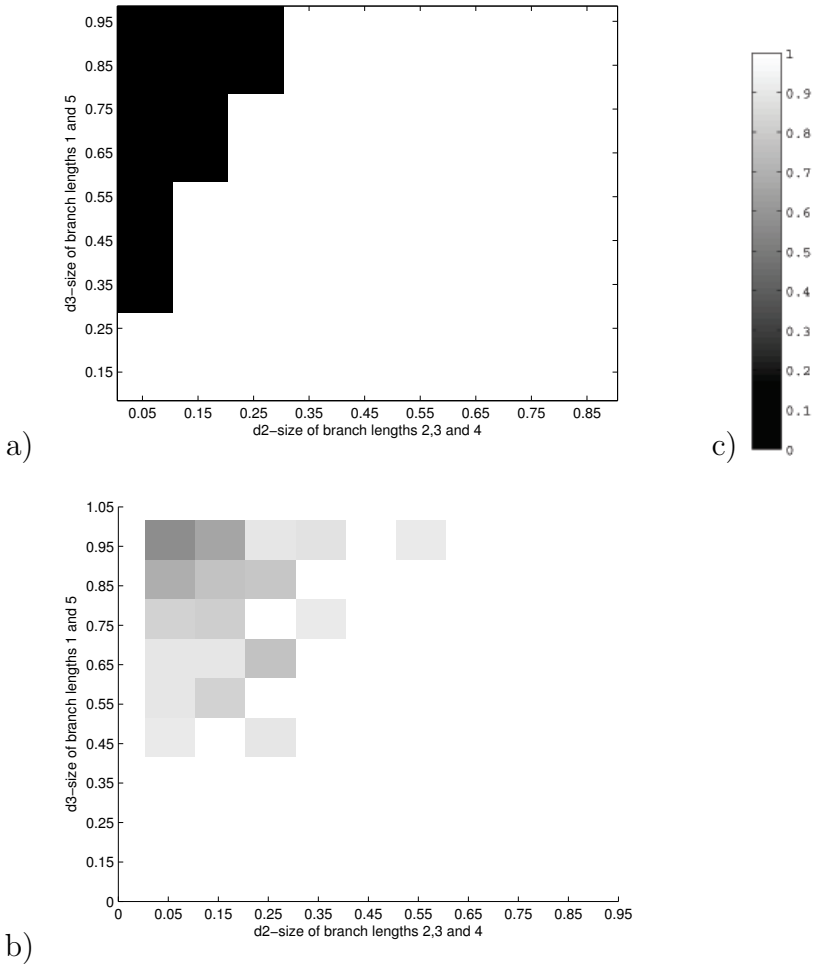


Figure 2.5: Posterior probabilities estimated with AIS

As in Figure 2.4, Subfigures a) and b) show the posterior probability of the correct tree topology for different branch length configurations. The results were obtained with annealed importance sampling rather than MCMC and show an average over five DNA sequence alignments independently generated for each branch length configuration. These configurations are determined by the values of $d2$ and $d3$, as defined in Figure 2.1. In each subfigure, the horizontal axis refers to $d2$, and the vertical axis refers to $d3$. The grey shading indicates the value of the inferred posterior probability, as indicated in the legend on the right, ranging from 0 (black) to 1 (white). Subfigure a) shows the results obtained for Model $\mathcal{H}_0$, represented in Figure 2.2. Subfigure b) shows the results obtained for Model $\mathcal{H}_1$, represented in Figure 2.3. Like in Figure 2.4, it is clearly observed that Model $\mathcal{H}_0$, which is shown in Subfigure a), leads to the systematic prediction of the wrong tree topology in the Felsenstein zone.

the marginal posterior probabilities $P(S_t|\mathcal{D})$ of the three possible tree topologies $S_t \in \{\Psi_1, \Psi_2, \Psi_3\}$[6]. The results were similar to those discussed in the previous section, with a clear systematic failure of PFHMM and DMCP in the Felsenstein zone. This failure was avoided when using PHMM. A specific example is presented in Figure 2.6, which shows the posterior probabilities $P(S_t|\mathcal{D})$ for a DNA sequence alignment $\mathcal{D}$ generated from the tree in Figure 2.1 with a branch length configuration $d2 = 0.15, d3 = 0.85$. A comparison with Figures 2.4 and 2.5 shows that this branch length configuration lies clearly in the Felsenstein zone. In support of our conjecture, both PFHMM and DMCP systematically show high posterior probabilities $P(S_t|\mathcal{D})$ close to 1 for a wrong tree topology throughout the whole DNA sequence alignment. Incidentally, in the high-scoring tree topology the two long non adjacent branches $d3$ in Figure 2.1 are grouped together, suggesting that PFHMM and DMCP suffer from the same long branch attraction as the method of maximum parsimony (Felsenstein, 1978b). There are no problems with PHMM, which consistently scored high posterior probabilities $P(S_t|\mathcal{D})$ close to 1 for the correct tree topology throughout the whole alignment.

2.12.3 Improving the phylogenetic factorial HMM

Our study described in the previous sections has revealed that the phylogenetic factorial HMM (PFHMM) of Husmeier (2005) is susceptible to systematically predicting spurious topology changes in the Felsenstein zone. The objective of the present section is to describe a modification of the PFHMM that avoids this shortcoming. A probabilistic graphical model representation of the PFHMM of Husmeier (2005) is shown in Figure 2.7 a. The model is essentially based on Model $\mathcal{H}_0$ of Figure 2.2 in that separate branch length vectors are associated with different sites of the alignment. This allows the branch lengths to be integrated out analytically, as described in Section 2.5, resulting in the simplified model depicted in Figure 2.7 b. The modified PFHMM is shown in Figure 2.7 c. Akin to Model $\mathcal{H}_1$ of Figure 2.3, a common vector of branch lengths is shared by all sites in the alignment[7]. The rate states $R_t \in \{\rho_1, \ldots, \rho_{k'}\}$, which in the original PFHMM of Husmeier (2005) are associated with the hyperparameter ρ of the

[6] These tree topologies are $\Psi_1 = (1,2,(3,4)), \Psi_2 = (1,3,(2,4))$, and $\Psi_3 = (1,4,(2,3))$, where the numbers refer to the four taxa.

[7] More accurately, there are three vectors of branch lengths $\mathbf{w}_k$, $k \in \{1,2,3\}$, associated with the three different tree topologies. This can be modelled as a common vector composed of three sub-vectors, where the state variable S_t indicates which of these subvectors applies to site t.

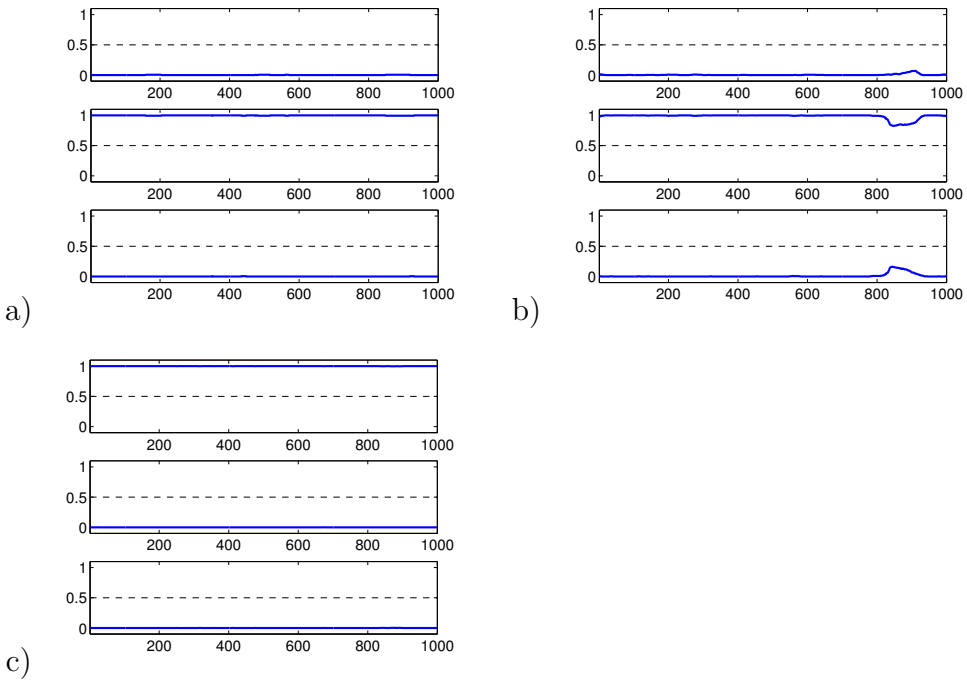


Figure 2.6: Failure of PFHMM and DMCP in the Felsenstein zone

Each figure shows a plot of the marginal posterior probability $P(S_t|\mathcal{D})$ (vertical axes) of the three possible tree topologies $S_t \in \{\Psi_1, \Psi_2, \Psi_3\}$ for the 4-taxa tree of Figure 2.1, plotted against the position t in the DNA sequence alignment (horizontal axes). Each subfigure consists of three panels, where the top panel corresponds to the true tree topology, from which the data were generated. The middle panel corresponds to a wrong tree topology, in which the two long branches $d3$ in Figure 2.1 are grouped together (long branch attraction). The bottom panel corresponds to another wrong tree topology. The three subfigures show the results obtained for the three recombination detection methods investigated: DCMP (Subfigure a), PFHMM (Subfigure b), and PHMM (Subfigure c). PHMM predicts high posterior probabilities $P(S_t|\mathcal{D})$ close to 1 for the true tree topology throughout the whole sequence alignment. However, both PFHMM and DMCP systematically show high posterior probabilities $P(S_t|\mathcal{D})$ close to 1 for the wrong tree topology in which the two long non-adjacent branches are joined.

prior distribution on the branch lengths, equation (2.9), are now associated with a global scaling factor by which the vector of branch lengths is multiplied. The hidden state sequences, **S** and **R**, and the model parameters are sampled from the posterior distribution with a Gibbs sampling procedure:

$$\mathbf{S}^{(i+1)} \sim P\left(\cdot|\mathbf{R}^{(i)},\nu_S^{(i)},\nu_R^{(i)},\mathbf{w}^{(i)},\mathcal{D}\right) \tag{2.44}$$

$$\mathbf{R}^{(i+1)} \sim P\left(\cdot|\mathbf{S}^{(i+1)},\nu_S^{(i)},\nu_R^{(i)},\mathbf{w}^{(i)},\mathcal{D}\right) \tag{2.45}$$

$$\nu_S^{(i+1)} \sim P\left(\cdot|\mathbf{S}^{(i+1)},\mathbf{R}^{(i+1)},\nu_R^{(i)},\mathbf{w}^{(i)},\mathcal{D}\right) \tag{2.46}$$

$$\nu_R^{(i+1)} \sim P\left(\cdot|\mathbf{S}^{(i+1)},\mathbf{R}^{(i+1)},\nu_S^{(i+1)},\mathbf{w}^{(i)},\mathcal{D}\right) \tag{2.47}$$

$$\mathbf{w}^{(i+1)} \sim P\left(\cdot|S^{(i+1)},\mathbf{R}^{(i+1)},\nu_S^{(i+1)},\nu_R^{(i+1)},\mathcal{D}\right) \tag{2.48}$$

where the superscript i denotes the iteration number. The first four steps are identical to those in Husmeier (2005): The hidden state sequences **S** and **R** are sampled with the stochastic forward-backward algorithm of Boys *et al.* (2000); the transition probabilities ν_S and ν_R, are sampled from beta distributions whose sufficient statistics are determined by **S** and **R**. The new aspect of our algorithm is the sampling of the branch length vector **w**. Since there is no closed-form expression for the distribution on the right-hand side of equation (2.48), we resort to a Metropolis-Hastings-within-Gibbs procedure. Note that $\mathbf{w}=(\mathbf{w}_1,\mathbf{w}_2,\mathbf{w}_3)$ is composed of three subvectors $\mathbf{w}_k$, $k\in\{1,2,3\}$, associated with the three tree topologies represented by the hidden state $S_t\in\{\Psi_1,\Psi_2,\Psi_3\}$. To ensure that the model is identifiable, we constrain the L1-norm of the branch length vectors to be equal to one: $||\mathbf{w}_k||_1=1$, $k\in\{1,2,3\}$; recall that the scaling of the branch lengths is effected by multiplication with a factor defined by the hidden states, $R_t\in\{\rho_1,\ldots,\rho_{K'}\}$. This constraint, as well as the positivity constraint $w_{ki}\geq 0$, is automatically guaranteed when proposing new branch length vectors $\mathbf{w}_k^*$ from a Dirichlet distribution:

$$Q(\mathbf{w}_k^*|\mathbf{w}_k) \propto \prod_i [w_{ki}^*]^{\alpha w_{ki}-1} \tag{2.49}$$

whose mean and variance are given by

$$\mathrm{E}[w_{ki}^*|w_{ki}] = w_{ki}; \quad \mathrm{Var}[w_{ki}^*|w_{ki}] = \frac{w_{ki}(1-w_{ki})}{\alpha+1} \tag{2.50}$$

Hence, the mean of the proposal distribution is equal to the current branch length, while the variance depends on a scaling parameter α. In our simulations, α was automatically adjusted in the burn-in phase to achieve an average acceptance

probability between 30% and 70%. The proposed vector of branch lengths $\mathbf{w}^*$ was accepted or rejected according to the standard Metropolis-Hastings criterion (Hastings, 1970), with the following acceptance probability:

$$A = \min\left\{1, \frac{L(\mathbf{w}_k^*)P(\mathbf{w}_k^*)Q(\mathbf{w}_k|\mathbf{w}_k^*)}{L(\mathbf{w}_k)P(\mathbf{w}_k)Q(\mathbf{w}_k^*|\mathbf{w}_k)}\right\} \tag{2.51}$$

where the proposal distribution $Q(\mathbf{w}_k^*|\mathbf{w}_k)$ is defined in equation (3.2.1), the prior distribution $P(\mathbf{w}_k)$ was chosen as defined in equation (2.9), with a fixed hyperparameter $\rho = 1$, and the likelihood $L(\mathbf{w}_k)$ depends on the hidden state sequences $\mathbf{S}$ and $\mathbf{R}$ as follows:

$$L(\mathbf{w}_k) = \prod_{t|S_t=\Psi_k} P(\mathbf{y}_t|R_t\mathbf{w}_k, \Psi_k, \theta) \tag{2.52}$$

where the expression in the argument of the product is given by equation (2.7). The details of the Gibbs sampling scheme used in our simulations are summarized in the appendix in section A.11.

2.12.4 Simulation details

We tested the improved PFHMM on the two types of synthetic DNA sequence alignments described in Section 3.3. The homogeneous DNA sequence alignments were the same as those used in the previous studies. The DNA sequence alignment with the mosaic structure was generated as described in Section 3.3, setting $d2 = d3 = 0.25$ for the flanking segments, and $d2 = 0.15, d3 = 0.85$ for the central segment. Hence, the branch length configuration corresponding to the central segment lies clearly in the Felsenstein zone; compare with Figures 2.4 and 2.5. Note that the DNA sequence alignment does not contain any change of the tree topology, though. For both the original PFHMM of Husmeier (2005) and the improved PFHMM we sampled the state sequences $\mathbf{S}$ from the posterior distribution with MCMC, monitoring convergence with the diagnostic test based on potential scale reduction factors (Gelman and Rubin, 1992); the details were given in Section 2.10.3. Note that both the original and the improved PFHMM need as input a set of fixed nucleotide substitution rates, corresponding to the hyperparameter ρ in equation (2.9). These values, which are associated with the rate states $\mathbf{R}$, were selected as follows: $\rho \in \{0.05, 0.1, 0.5, 1, 2, 4, 6, 8\}$.

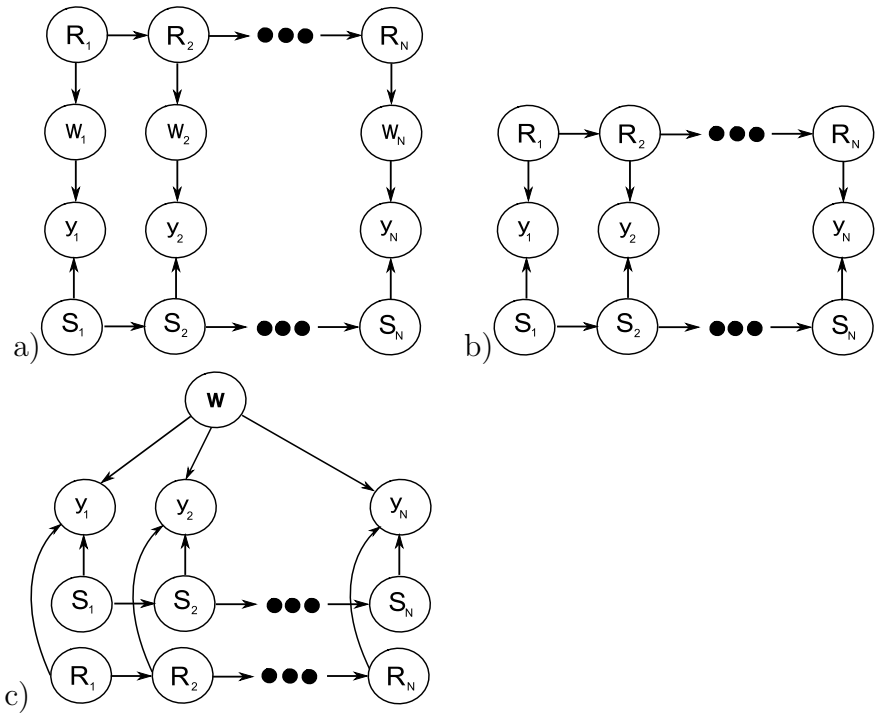
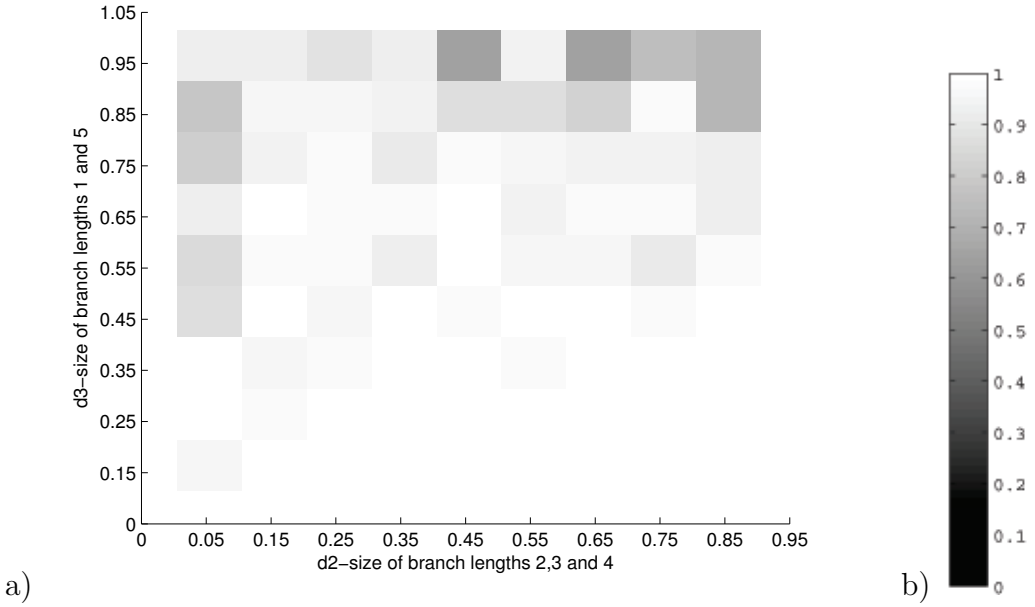


Figure 2.7: Graphical model of the PFHMM and the improved PFHMM

Subfigure a) shows the probabilistic graphical model representation of the phylogenetic factorial HMM of Husmeier (2005). The $\mathbf{y}_t$'s represent the columns in the DNA sequence alignment, where the subscript $t = 1, \ldots, N$ indicates the site in the alignment. Each site t is associated with a hidden state S_t that defines the tree topology, a vector of branch lengths $\mathbf{w}_t$, and a second hidden state R_t that defines the hyperparameter of the prior distribution on the branch lengths, as defined in equation (2.9). Both hidden states S_t and R_t have a Markovian dependence structure. The chosen form of the model allows the branch lengths to be integrated out analytically, as described in Section 2.5. This results in the simplified model depicted in Subfigure b). Note that this model is a phylogenetic factorial HMM, where one type of hidden states $(S_1, \ldots, S_N)$ defines the tree topology, and the other type of hidden states $(R_1, \ldots, R_N)$ defines the average amount of mutational divergence. Hence, the model presented here is a generalization of the model shown in Figure 2.2 so as to allow for recombination and rate heterogeneity. Subfigure c) shows the probabilistic graphical model representation of the improved phylogenetic factorial HMM proposed in the present article. The model is similar to the one presented in the previous subfigures with the difference that a common branch length vector $\mathbf{w}$ is shared among all sites. This is a generalization of the standard phylogenetic model of Figure 2.3 that allows for recombination and rate heterogeneity.



a) b)

Figure 2.8: Results of the improved PFHMM

The figure shows, for different branch length configurations $[d2, d3]$, as defined in Figure 2.1, the posterior probability $\overline{P}(S = \Psi_{true}|\mathcal{D})$, defined in equation (2.53). The horizontal axis represents $d2$, and the vertical axis represents $d3$. Probabilities are represented by a grey shading, ranging from white (1) to black (0), as indicated by the legend on the right. The figure shows that as a consequence of the modification of the PFHMM, described in Section 2.12.3, the systematic failure in the Felsenstein zone, which was found in Subfigure a) of Figures 2.4 and 2.5, is avoided. These results are the averaging over 2 independent simulations.

2.12.5 Simulation results

Figure 2.8 shows the results obtained with the improved PFHMM on the homogeneous DNA sequence alignment. The figure shows, for different branch length configurations $[d2, d3]$, the average posterior probability of the correct tree topology Ψ_{true}, averaged over all positions in the alignment:

$$\overline{P}(S = \Psi_{true}|\mathcal{D}) \;=\; \frac{1}{N}\sum_{t=1}^{N} P(S_t = \Psi_{true}|\mathcal{D}) \tag{2.53}$$

It is clearly seen that the failure in the Felsenstein zone is avoided, and that $\overline{P}(S = \Psi_{true}|\mathcal{D})$ is consistently greater than 0.5 (and close to 1 in most cases).

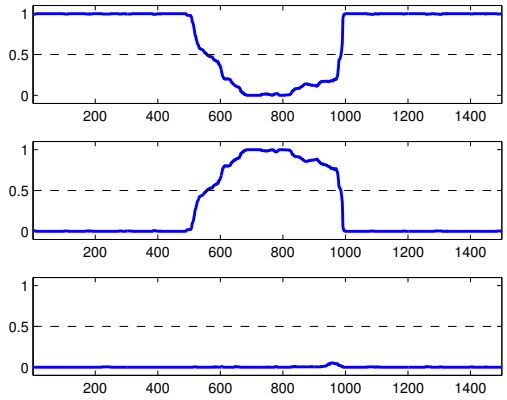
Figure 2.9 shows the results obtained on the DNA sequence alignment with the mosaic structure. Both subfigures show a plot of the predicted marginal posterior probabilities $P(S_t = \Psi_i|\mathcal{D})$, for the three possible tree topologies $i \in \{1, 2, 3\}$,

plotted against the position t in the alignment. The left subfigure shows the prediction obtained with the original PFHMM of Husmeier (2005). There is a clear transition into a different tree topology in the central region, where the branch length configuration $[d2, d3]$ lies in the Felsenstein zone. This confirms our conjecture that PFHMM is susceptible to the prediction of spurious topology changes. The right panel shows the prediction made with the improved PFHMM averaged over 5 independent simulations. The posterior probability for the correct tree topology, $P(S_t = \Psi_{true}|\mathcal{D})$, is consistently close to 1, indicating that the prediction of spurious topology changes is avoided.

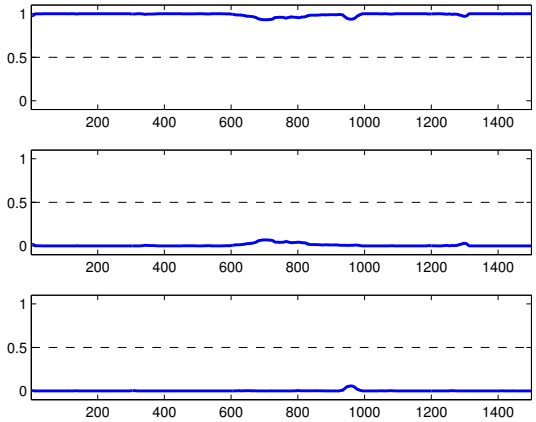
2.13 Discussion

In this chapter we have investigated a possible shortcoming of three recent Bayesian methods for detecting recombination in DNA sequence alignments: the multiple change-point (MCP) model of Suchard *et al.* (2003), the dual multiple change-point (DMCP) model of Minin *et al.* (2005), and the phylogenetic factorial hidden Markov model (PFHMM) of Husmeier (2005). All three models assume separate branch lengths for different sites, which allows the branch lengths to be integrated out analytically. This reduces the computational complexity of the Bayesian inference scheme, which can now be formulated in terms of posterior distributions of the tree topologies and the nucleotide substitution parameters only. This makes the approach quite popular, and it has been applied in more recent works; see Lehrach (2008) and Lehrach and Husmeier (2009).

Note that the model of site-independent branch lengths, as expressed in eq. (2.16), was first introduced by Tuffley and Steel (1997), where it was called the ‘no-common-mechanism’ model. In combination with the prior independence of the branch length components, expressed in eq. (2.9), the vector of branch lengths can be integrated out in the likelihood, as shown by Suchard *et al.* (2003), and discussed in Section 2.5. However, in the no-common-mechanism model, the branch lengths are incidental rather than structural parameters. As discussed in Goldman (1990), this implies that maximum likelihood is no longer guaranteed to provide a consistent estimator. In fact, Tuffley and Steel (1997) showed that under certain regularity conditions, maximum parsimony and maximum likelihood with no common mechanisms are equivalent. This suggests that maximum likelihood with no common mechanisms will be susceptible to the prediction of wrong tree



a)



b)

Figure 2.9: Mosaic DNA sequence alignment

The figure shows the predictions obtained with the original PFHMM of Husmeier (2005) versus the improved PFHMM proposed in this chapter. Both models were applied to a synthetic DNA sequence alignment with mosaic structure, where the branch length configuration in the central segment lies in the Felsenstein zone; see the description in Section 3.3. Each panel shows a plot of the predicted marginal posterior probabilities $P(S_t = \Psi_i|\mathcal{D})$ for the three possible tree topologies $i \in \{1,2,3\}$, where the true topology corresponds to the panel in the top. The vertical axes show the marginal posterior probabilities, while the horizontal axes represent the site t in the alignment. The original PFHMM, shown in Subfigure a), predicts a spurious topology change in the central segment, which is avoided with the improved PFHMM, shown in Subfigure b). Subfigure b) is an average over 5 independent simulations showing that the spurious topology change is avoided consistently.

topologies for certain branch length configurations (long branch attraction). To confirm this hypothesis, we have generated synthetic DNA sequence alignments with the HKY nucleotide substitution model (Hasegawa *et al.*, 1985) in the vein of Felsenstein's seminal study for demonstrating the inconsistency of maximum parsimony Felsenstein (1978b), and we have estimated the marginal posterior probability for the tree topology in two different ways: using an inter-model approach, in which tree topologies are sampled from the posterior distribution with MCMC; and applying an intra-model approach, in which the marginal likelihood is estimated with annealed importance sampling. Both studies consistently reveal that as a consequence of the separate site-dependent branch lengths, the mode of the posterior distribution is systematically shifted to a wrong tree topology whenever the branch length configuration of the data-generating tree falls into the Felsenstein zone. The inferred tree topology with the highest posterior probability is the one in which the long branches are grouped together. This finding suggests that as a consequence of the aforementioned independence assumption (i.e., the "no-common-mechanism" model of separate site-dependent branch lengths), the resulting model suffers from the same inconsistency (long-branch attraction) as the method of maximum parsimony. We have further confirmed this conjecture by applying the recombination detection methods DMCP and PFHMM to the DNA sequence alignments generated in our study, using the authors' programs. Again, we found a systematic failure in the Felsenstein zone, where consistently the wrong tree topology was inferred. This suggests that these recombination detection methods are susceptible to predicting spurious recombination events whenever branch-length configurations happen to fall near the boundary of the Felsenstein zone.

We have concluded our study with a demonstration of how the PFHMM can be improved to avoid this shortcoming. In principle this can be achieved by removing the site-independence assumption for the branch lengths. As a consequence, however, the analytic integration over the branch lengths is no longer tractable, which requires them to be sampled approximately from the posterior distribution with MCMC. To avoid an identifiability problem resulting from the fact that the global scaling of the branch lengths (defined by one of the two types of hidden states) is an additional independent parameter of the model, we have imposed a normalization constraint on the branch lengths, which can easily be effected by the choice of a suitable proposal distribution in the MCMC scheme. We have tested

the proposed method on the same DNA sequence alignments as for the other models, and found that it succeeded in avoiding the failure in the Felsenstein zone.

Note that in the proposed phylogenetic PFHMM, each hidden state is associated with a distinct tree topology. The number of tree topologies increases super-exponentially with the number of taxa; for this reason, we have applied our model to DNA sequence alignments of four sequences only, as in Husmeier and McGuire (2003). There are various heuristic simplifications one could adopt in order to apply the method to sequence alignments with more than four taxa. One method would be to apply a preliminary phylogenetic analysis to consecutive subsets of the DNA sequence alignment, effected for instance in the way described in Husmeier *et al.* (2005b). The phylogenetic FHMM would then include only those topology states that match one of the tree topologies inferred in the preliminary analysis. Another method would be to proceed in the way described in Minin *et al.* (2005). Here, the assumption is that we are given a sequence alignment composed of N-1 nonrecombinant and 1 putative recombinant strain. Additionally, it is assumed that the tree of the N-1 nonrecombinant sequences is known or that it can easily be inferred. The states of the phylogenetic FHMM are restricted to the set of those tree topologies that are obtained by adding a new leaf node to any branch in the fixed parental tree with $N-1$ nonrecombinant taxa. Both of these heuristic simplifications substantially restrict the set of permissible tree topologies, thereby rendering the application of the phylogenetic FHMM to larger alignments viable. Note, though, that in principle these restrictions are not necessary. Our method could in principle be implemented with a transdimensional MCMC scheme using reversible jumps associated with the birth and death of topology states, where each birth creates a new tree topology derived from the adjacent topology by some local modification, e.g. using nearest neighbour interchange. However, the computational costs of such an approach would be huge, and it would pose a challenging problem for novel high-performance distributed computing techniques.

It has been pointed out by one of the referees that our work is closely related to the work of Huelsenbeck *et al.* (2008). Like in our chapter, the authors investigate a Bayesian implementation of the no-common-mechanism model, and they empirically demonstrate that this model is not consistent and shows a systematic failure in the Felsenstein zone. There are various ways in which our study

complements this work. Firstly, Huelsenbeck *et al.* (2008) use a fully symmetric nucleotide substitution model that makes no distinction between any character states (the Jukes-Cantor model). For this model, Tuffley and Steel (1997) showed that maximum parsimony and maximum likelihood with no common mechanism are equivalent in the sense that both choose the same tree. Hence, the work of Huelsenbeck *et al.* (2008) can be seen as an empirical corroboration of the theoretical findings in Tuffley and Steel (1997). Our study complements this work by using the HKY model (Hasegawa *et al.*, 1985) as a more general and more widely applied nucleotide substitution model, for which no theoretical proof was given in Tuffley and Steel (1997). Secondly, Huelsenbeck *et al.* (2008) use fixed parameters for the prior distribution on the branch lengths and find that these parameters "play an inordinately strong role in determining the probabilities of the trees". In our work on the homogeneous DNA sequence alignment, we use a hierarchical Bayesian model with an extra layer (a hyperprior) – see Figures 2.2 and 2.3 – and infer the parameters of the prior from the data. In our work on the DNA sequence alignment with mosaic structure, we use a phylogenetic FHMM, in which the parameters of the prior distribution are associated with different hidden states. The assignment of these hidden states to sites is inferred from the data. Thirdly, one has to appreciate that there is no sufficient criterion to prove that an MCMC simulation has converged. For this reason computing the posterior probabilities of tree topologies with an alternative paradigm, as we do in our intra-model approach based on annealed importance sampling, offers an independent corroboration of the findings. Fourthly and most importantly, however, there has been a completely different focus of our work. The motivation for the work of Huelsenbeck *et al.* (2008) has been the development of a new Bayesian MCMC scheme for learning tree topologies from sequence alignments that are adequately described by a single tree. The focus of our study is the prediction of recombination and mosaic structures in DNA sequence alignments, and it has been motivated by three recent detection methods that are based on the no-common-mechanism model. These models are more flexible than the single-tree model investigated by Huelsenbeck *et al.* (2008). In particular, they allow for breakpoints in the DNA sequence alignment at which the tree topology may change. While this mechanism provides the extra flexibility required for dealing with recombination, we have shown that in combination with site-independent branch lengths (the no-common-mechanism of Tuffley and Steel (1997)), the re-

sulting model becomes susceptible to predicting spurious topology changes and recombination breakpoints.

2.14 Future work

The phylogenetic FHMM proposed in Section 2.12.3 of the chapter provides a trade-off between two extreme scenarios: the homogeneous model, which employs the same branch lengths for the whole alignment, and the no-common-mechanism model. The first approach is too restrictive. In the second approach, the branch lengths are incidental rather than structural parameters, resulting in the inconsistency problems discussed in the present chapter. The proposed phylogenetic FHMM contains a hidden factor by which the branch lengths are rescaled. This scaling factor is site dependent via its association with a hidden state of the FHMM. Since the number of hidden states is finite, and each hidden state can be revisited repeatedly when traversing along the alignment, all parameters of the model are structural (rather than incidental). In this way, the consistency of our model is guaranteed. However, while our model is appropriate to incorporate the effects of rate heterogeneity, it is too restrictive when dealing with certain recombination events that do not induce a tree topology change. This can happen when, in the coalescence tree, recombinant lineages coalesce before merging with any other lineage (Wiuf *et al.*, 2001). In certain scenarios, discussed in Wiuf *et al.* (2001), this can result in a more complex change of the branch lengths than can be modelled by a global rescaling. One way to proceed would be the following modification of our model. Rather than associating the second hidden state with a global scaling factor, we could associate it with a separate vector of branch lengths. In this model there would no longer be a common branch length vector, but the branch lengths would be site-dependent, as in the no-common-mechanism model. The substantial difference from the no-common-mechanism model would be that in the new model the site dependence is effected indirectly via a hidden state. Since the number of hidden states is finite, and the hidden states can be revisited (at least as long as all transition probabilities are non-zero), the new model contains structural rather than incidental parameters. In this way, its consistency is guaranteed. Note, however, that this model is more complex than the one proposed in this chapter. In particular, it will require the number of hidden states and their associated parameters to be properly inferred

from the data rather than chosen in advance. This calls for the development of a trans-dimensional MCMC scheme with RJMCMC (Green, 1995), as applied in the studies by Suchard *et al.* (2003), Minin *et al.* (2005), Lehrach (2008) and Lehrach and Husmeier (2009). We believe that this would be an important and stimulating topic for future research.

When extending the phylogenetic FHMM along the line discussed in the previous paragraph, one has to decide on the appropriate form of the prior distribution on tree topologies. It is a common approach in Bayesian analysis to use a uniform prior distribution. The intention is to reflect our prior level of ignorance, as especially promoted by the school of öbjective Bayesianism". The question, then, is what exactly it is that we are ignorant about. A prior distribution that is uniform over tree topologies is not uniform over labelled histories or clade formations, where the latter inconsistency has been used to (erroneously!) question the validity of the Bayesian approach per se (Pickett and Randle, 2005). As pointed out by Velasco (2008), the ignorance should be expressed in terms of the physical processes that generate the entities of interest. A phylogenetic tree is the result of the biological process of common ancestry and descent with modification, which can be modelled by a Yule random branching process (forward in time) or a coalescence process (backward in time). Kingman (1982) and Thompson (1975) showed that under certain regularity conditions, the Yule birth process, the Yule birth-death process and the coalescence process lead to the same distribution. This distribution is uniform over labelled histories (Edwards, 1970), which induces a prior distribution on tree topologies that is no longer uniform. In particular, Velasco (2008) showed that a tree topology that is more balanced (as opposed to pectinate) is consistent with more labelled histories and, consequently, has a higher prior probability. An early application of this approach can be found in Yang and Rannala (1997). However, the computational costs were found to be huge – about two orders of magnitude larger than those of the competing method of Larget and Simon (1999). It therefore will pose a substantial computational challenge for future work to render the approach based on labelled histories viable in the context of the model proposed in the present chapter.

Chapter 3

An improved model to distinguish between global and within-codon rate variation

This chapter follows chapter 2 as an improvement. An intrinsic failure in the model of chapter 2 to correctly distinguish between the short range rate heterogeneity on the codon level and long range rate heterogeneity is investigated. The model is improved to fit both features of the data.

3.1 Introduction

DNA sequence alignments are usually not homogeneous. Mosaic structures may result as a consequence of recombination or rate heterogeneity. Interspecific recombination, in which DNA subsequences are transferred between different (typically viral or bacterial) species may result in a change of the topology of the underlying phylogenetic tree. Rate heterogeneity corresponds to a change of the nucleotide substitution rate. Two Bayesian methods for simultaneously detecting recombination and rate heterogeneity in DNA sequence alignments are the dual multiple change-point model (DMCP) of Minin *et al.* (2005), and the phylogenetic factorial hidden Markov model (PFHMM) of Husmeier (2005) and Lehrach and Husmeier (2009). The idea underlying the DMCP is to segment the DNA sequence alignment by the insertion of change-points, and to infer different phylogenetic trees and nucleotide substitution rates for the separate segments thus obtained. Two separate change-point processes associated with the tree topology

and the nucleotide substitution rate are employed. Inference is carried out in a Bayesian way with reversible jump (RJ) Markov chain Monte Carlo (MCMC). Of particular interest are the number and locations of the change-points, which mark putative recombination break-points and regions putatively under different selective pressures. A related modelling paradigm is provided by the PFHMM, where two *a priori* independent hidden Markov chains are introduced, whose states represent the tree topology and nucleotide substitution rate, respectively. While the earlier work of Husmeier (2005) kept the number of hidden states fixed, Lehrach and Husmeier (2009) generalised the inference procedure with RJMCMC and showed that this framework subsumes the DMCP as a special case. This model has recently been extended to larger numbers of species Webb *et al.* (2009).

Common to all these models are two simplifications. First, the no-common mechanism model of Tuffley and Steel (1997) is introduced, which assumes separate branch lengths for each site in the DNA sequence alignment. Second, there is no distinction between regional and within-codon rate heterogeneity. Following Suchard *et al.* (2003), the first assumption was introduced with the objective to reduce the computational complexity of the inference scheme. The no-common-mechanism model allows the branch lengths to be integrated out analytically. This is convenient, as the marginal likelihood of the tree topology, the nucleotide substitution rate, and further parameters of the nucleotide substitution model (like the transition- transversion ratio) can be computed in closed from. In this way, the computational complexity of sampling break-points (DMCP) or hidden state sequences (PFHMM) from the posterior distribution with MCMC is substantially reduced. However, in the no-common-mechanism model the branch lengths are incidental rather than structural parameters. As we discussed in chapter 2 and presented in Husmeier and Mantzaris (2008), this implies that maximum likelihood no longer provides a consistent estimator, and that the method systematically infers the wrong tree topology in the Felsenstein zone defined in Felsenstein (1978b). The second simplification does not distinguish between two different types of rate heterogeneity: (1) a regional effect, where larger consecutive segments of the DNA sequence alignment might be differently evolved, e.g. as a consequence of changes of the selective pressure; (2) and a codon effect, where the third codon position shows more variation than the first or the second. Not allowing for this difference and treating both sources of rate heterogeneity on an equal footing implies the risk that subtle regional effects might be obscured by

the short-range codon effect, as discussed in Lehrach and Husmeier (2009). The latter effect is of no biological interest, though, as it only represents the signature of the genetic code.

In the present work, we address this issue and develop a model that properly distinguishes between these two effects. Our work is based on the model we introduced in chapter 2 and presented in Husmeier and Mantzaris (2008). We modify this approach so as to explicitly take the signature of the genetic code into account. In this way, the within-codon effect of rate heterogeneity is imposed on the model *a priori*, which makes it easier to learn the biologically more interesting effect of regional rate heterogeneity *a posteriori*. The work of this chapter has already been published in Mantzaris and Husmeier (2009).

3.2 Methodology

3.2.1 Modelling recombination and rate heterogeneity with a phylogenetic FHMM

Consider an alignment $\mathcal{D}$ of m DNA sequences, N nucleotides long. Let each column in the alignment be represented by $\mathbf{y}_t$, where the subscript t represents the site, $1 \leq t \leq N$. Hence $\mathbf{y}_t$ is an m-dimensional column vector containing the nucleotides at the tth site of the alignment, and $\mathcal{D} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$. Given a probabilistic model of nucleotide substitutions based on a homogeneous Markov chain with instantaneous rate matrix $\mathbf{Q}$, a phylogenetic tree topology S, and a vector of branch lengths $\mathbf{w}$, the probability of each column $\mathbf{y}_t$, $P(\mathbf{y}_t|S,\mathbf{w},\theta)$, can be computed, as e.g. discussed in Felsenstein (1981). Here, θ denotes a (vector) of free nucleotide substitution parameters extracted from $\mathbf{Q}$. For instance, for the HKY85 model of Hasegawa *et al.* (1985), we have $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$, with $\pi_i \in [0,1]$ and $\sum_i \pi_i = 1$, is a vector of nucleotide equilibrium frequencies, and $\alpha, \beta \geq 0$ are separate nucleotide substitution rates for transitions and transversions. For identifiability between $\mathbf{w}$ and $\mathbf{Q}$, the constraint $\sum_i Q_{ii}\pi_i = -1$ is commonly introduced, which allows the branch lengths to be interpreted as expected numbers of mutations per site (see, e.g., Minin *et al.* (2005)). The normalisation constraint on π further reduces the number of free parameters by one, so that without loss of generality we have $\theta = (\pi_A, \pi_C, \pi_G, \zeta)$, where $\zeta = \alpha/(2\beta) \geq 0$ is the transition-transversion ratio. In what follows, we do not make the dependence

on θ explicit in our notation.

We simultaneously model recombination and rate heterogeneity with a phylogenetic FHMM, as originally proposed in Husmeier (2005), with the modification discussed in chapter 2 (presented in Husmeier and Mantzaris (2008)). A hidden variable $S_t \in \{\tau_1, \ldots, \tau_K\}$ is introduced, which represents one out of K possible tree topologies τ_i at site t. To allow for correlations between nearby sites – while keeping the computational complexity limited – a Markovian dependence structure is introduced: $P(\mathbf{S}) = P(S_1, \ldots, S_N) = \prod_{t=2}^{N} P(S_t|S_{t-1})P(S_1)$. Following Felsenstein and Churchill (1996), the transition probabilities are defined as

$$P(S_t|S_{t-1}, \nu_S) = \nu_S^{\delta(S_t, S_{t-1})} \left(\frac{1-\nu_S}{K-1}\right)^{[1-\delta(S_t, S_{t-1})]} \tag{3.1}$$

where $\delta(S_t, S_{t-1})$ denotes the Kronecker delta symbol, which is 1 when $S_t = S_{t-1}$, and 0 otherwise. The parameter ν_S denotes the probability of not changing the tree topology between adjacent sites. Associated with each tree topology τ_i is a vector of branch lengths, $\mathbf{w}_{\tau_i}$, which defines the probability of a column of nucleotides, $P(\mathbf{y}_t|S_t, \mathbf{w}_{S_t})$. The practical computation follows standard methodology based on the pruning algorithm Felsenstein (1981). For notational convenience we rewrite these *emission probabilities* as $P(\mathbf{y}_t|S_t, \mathbf{w})$, where $S_t \in \{\tau_1, \ldots, \tau_k\}$ determines which of the subvectors $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_K)$ is selected. To model rate heterogeneity, a second type of hidden states R_t is introduced. Correlations between adjacent sites are modelled again by a Markovian dependence structure: $P(\mathbf{R}) = P(R_1, \ldots, R_N) = \prod_{t=2}^{N} P(R_t|R_{t-1})P(R_1)$. The transition probabilities are defined as in (3.1):

$$P(R_t|R_{t-1}, \nu_R) = \nu_R^{\delta(R_t, R_{t-1})} \left(\frac{1-\nu_R}{\tilde{K}-1}\right)^{[1-\delta(R_t, R_{t-1})]} \tag{3.2}$$

where $\tilde{K}$ is the total number of different rate states. Each rate state is associated with a scaling parameter $R_t \in \rho = \{\rho_1, \ldots, \rho_{K'}\}$ by which the branch lengths are rescaled: $P(\mathbf{y}_t|S_t, \mathbf{w}) \rightarrow P(\mathbf{y}_t|S_t, R_t\mathbf{w})$. To ensure that the model is identifiable, we constrain the L1-norm of the branch length vectors to be equal to one: $||\mathbf{w}_k||_1 = 1$ for $k = 1, \ldots, K$. To complete the specification of the probabilistic model, we introduce prior probabilities on the transition parameters ν_S and ν_R, which are given conjugate beta distributions (which subsume the uniform distribution for the uninformative case). The initial state probabilities $P(S_1)$ and $P(R_1)$ are set to the uniform distribution, as in Husmeier and McGuire (2003). The prediction

of recombination break-points and rate heterogeneity is based on the marginal posterior probabilities

$$P(S_t|\mathcal{D}) = \sum_{S_1}\dots\sum_{S_{t-1}}\sum_{S_{t+1}}\dots\sum_{S_N} P(\mathbf{S}|\mathcal{D}) \quad (3.3)$$

$$P(R_t|\mathcal{D}) = \sum_{R_1}\dots\sum_{R_{t-1}}\sum_{R_{t+1}}\dots\sum_{R_N} P(\mathbf{R}|\mathcal{D}) \quad (3.4)$$

The distributions $P(\mathbf{S}|\mathcal{D})$ and $P(\mathbf{R}|\mathcal{D})$ are obtained by the marginalisation

$$P(\mathbf{S}|\mathcal{D}) = \sum_{\mathbf{R}}\int P(\mathbf{S},\mathbf{R},\nu_S,\nu_R,\mathbf{w}|\mathcal{D})d\nu_S d\nu_R d\mathbf{w} \quad (3.5)$$

$$P(\mathbf{R}|\mathcal{D}) = \sum_{\mathbf{S}}\int P(\mathbf{R},\mathbf{S},\nu_S,\nu_R,\mathbf{w}|\mathcal{D})d\nu_S d\nu_R d\mathbf{w} \quad (3.6)$$

where $P(\mathbf{S},\mathbf{R},\nu_S,\nu_R,\mathbf{w}|\mathcal{D}) \propto P(\mathcal{D},\mathbf{S},\mathbf{R},\nu_S,\nu_R,\mathbf{w}) = P(S_1)P(R_1)P(\nu_S)P(\nu_R)$ $\prod_{t=1}^{N} P(\mathbf{y}_t|S_t,R_t\mathbf{w})\prod_{t=2}^{N} P(S_t|S_{t-1},\nu_S)\prod_{t=2}^{N} P(R_t|R_{t-1},\nu_R)$. The respective integrations and summations are intractable and have to be numerically approximated with Markov chain Monte Carlo (MCMC): we sample from the joint posterior distribution $P(\mathbf{S},\mathbf{R},\nu_S,\nu_R,\mathbf{w}|\mathcal{D})$ and then marginalise with respect to the entities of interest. Sampling from the joint posterior distribution follows a Gibbs sampling procedure Casella and George (1992), where each parameter group is iteratively sampled separately conditional on the others. So if the superscript (i) denotes the ith sample of the Markov chain, we obtain the $(i+1)$th sample as follows:

$$\mathbf{S}^{(i+1)} \sim P(\cdot|\mathbf{R}^{(i)},\nu_S^{(i)},\nu_R^{(i)},\mathbf{w}^{(i)},\mathcal{D}) \quad (3.7)$$

$$\mathbf{R}^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)},\nu_S^{(i)},\nu_R^{(i)},\mathbf{w}^{(i)},\mathcal{D}) \quad (3.8)$$

$$\nu_S^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)},\mathbf{R}^{(i+1)},\nu_R^{(i)},\mathbf{w}^{(i)},\mathcal{D}) \quad (3.9)$$

$$\nu_R^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)},\mathbf{R}^{(i+1)},\nu_S^{(i+1)},\mathbf{w}^{(i)},\mathcal{D}) \quad (3.10)$$

$$\mathbf{w}^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)},\mathbf{R}^{(i+1)},\nu_S^{(i+1)},\nu_R^{(i+1)},\mathcal{D}) \quad (3.11)$$

The order of these sampling steps is arbitrary. Note that, in principle, the nucleotide substitution parameters θ should be included in the Gibbs scheme, as described in Husmeier and McGuire (2003). In practice, a fixation of θ at *a priori* estimated values makes little difference to the prediction of $P(S_t|\mathcal{D})$ and $P(R_t|\mathcal{D})$ and has the advantage of reduced computational costs. Changing the value of the parameters for the evolutionary model does not incur changes to the mosaic structure (changes in the break points) inferred for the ratestates or

the topologies along the sites of the alignment. A wrongly assigned parameter θ would alter the ratefactor values inferred for all sites, rather than segments, and since this applies to all the sites it does not remove the information needed to infer the break points which is the primary purpose of this chapter.

Sampling the hidden state sequences $\mathbf{S}$ and $\mathbf{R}$ in (3.7) and (3.8) is effected with the stochastic forward-backward algorithm of Boys *et al.* (2000). Sampling the transition probabilities ν_S and ν_R in (3.9) and (3.10) is straightforward due to the conjugacy of the beta distribution. Sampling the branch lengths in (3.11) cannot be effected from a closed-form distribution, and we have to resort to a Metropolis-Hastings-within-Gibbs scheme. Note that the branch lengths have to satisfy the constraint $||\mathbf{w}_k||_1 = 1$, $k = 1,\ldots,K$, as well as the positivity constraint $w_{ki} \geq 0$. This is automatically guaranteed when proposing new branch length vectors $\mathbf{w}_k^*$ from a Dirichlet distribution: $Q(\mathbf{w}_k^*|\mathbf{w}_k) \propto \prod_i [w_{ki}^*]^{\alpha w_{ki}-1}$, where α is a tuning parameter that can be adapted during burn-in to improve mixing. The acceptance probability for the proposed branch lengths is then given by the standard Metropolis-Hastings criterion Hastings (1970).

3.2.2 Distinguishing regional from within-codon rate heterogeneity

We improve the model described in the previous subsection, which is shown in chapter 2 (proposed in Husmeier and Mantzaris (2008)), in two respects. First, we adapt ρ and sample it along with $\mathbf{w}$ from the posterior distribution. The sampling procedure mirrors that for the branch lengths. It is done with MCMC as defined in eq.. The priors here are chosen to be uniform as all penalisations made it difficult for rates to be accepted which were large enough to compensate for the normalised codon vector applied to the model. To make the ratefactor notation explicit in the notation, we slightly change the definition of the rate state as $R_t \in \{1,\ldots,K'\}$ and rewrite: $P(\mathbf{y}_t|S_t, R_t\mathbf{w}) \rightarrow P(\mathbf{y}_t|S_t, \rho_{R_t}\mathbf{w})$. Second, we explicitly model codon-position-specific rate heterogeneity in a way similar to Felsenstein and Churchill (1996). This work applied an HMM to infer rate heterogeneity break points along sequence alignments. To this end, we introduce the indicator variable $I_t \in \{0,1,2,3\}$, where $I_t = 0$ indicates that the tth position of the alignment does not code for protein, and $I_t = i \in \{1,2,3\}$ indicates that site t is the ith position of a codon. Each of the four categories is associated with a positive factor

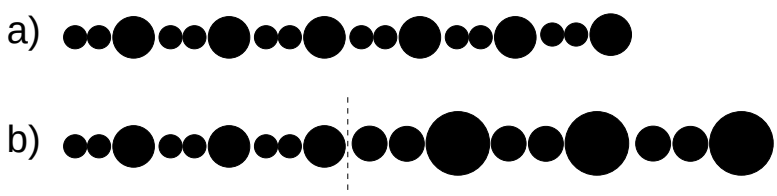


Figure 3.1: Illustration of regional versus within-codon rate heterogeneity. Each circle corresponds to a nucleotide in a DNA sequence, and the circle diameter symbolises the average nucleotide substitution rate at the respective position. The top panel (a) shows a "homogeneous" DNA sequence composed of six codons, where each third position is more diverged as a consequence of the nature of the genetic code. The bottom panel (b) shows a hypothetical DNA sequence subject to regional rate heterogeneity, where the second half on the right of the dashed vertical line constitutes a region that is more evolved. The sequences used in our simulation study were similar, but longer (1.5Kbp).

taken from $\lambda = (\lambda_0, \lambda_1, \lambda_2, \lambda_3)$, by which the branch lengths are modulated. The emission probabilities are thus given by $\tilde{P}(\mathbf{y}_t|S_t, R_t, I_t, \rho, \lambda, \mathbf{w}) := P(\mathbf{y}_t|S_t, \rho_{R_t}\lambda_{I_t}\mathbf{w})$, where $P(.)$ was defined below equation (3.1), and $\tilde{P}(.)$ makes the dependence on ρ and λ explicit. Note that as opposed to Felsenstein and Churchill (1996), we do not keep λ fixed, but sample it from the posterior distribution with MCMC. For identifiability we introduce the same constraint as for the branch lengths: $||\lambda||_1 = 1$, which is automatically guaranteed when proposing λ from a Dirichlet distribution. Hence, to sample ρ and λ from the posterior distribution $P(\mathbf{S}, \mathbf{R}, \nu_S, \nu_R, \rho, \lambda, \mathbf{w}|\mathcal{D})$, we have to add two Metropolis-Hastings-within-Gibbs steps akin to equation (3.11) to the Gibbs sampling procedure (3.7-3.11):

$$[\rho^{(i+1)}, \lambda^{(i+1)}] \sim P(\cdot|\mathbf{S}^{(i+1)}, \mathbf{R}^{(i+1)}, \nu_S^{(i+1)}, \nu_R^{(i+1)}, \mathbf{w}^{(i+1)}, \mathcal{D}) \qquad (3.12)$$

With all other parameters and hidden states fixed, we propose new values for ρ and λ, and accept or reject according to the Metropolis-Hastings criterion. As discussed above, we propose new values for λ from a Dirichlet distribution. New values for ρ are proposed from a uniform distribution (on the log scale), centred on the current values. The dispersal parameters of the proposal distributions can be adjusted during the burn-in phase using standard criteria.

3.3 Data

To assess the performance of the method, we tested it on synthetic DNA sequence alignments; this has the advantage that we have a known gold-standard. For a realistic simulation, we generated sequence alignments with Seq-Gen, developed by Rambaut and Grassly. This software package is widely used for Monte Carlo simulations of molecular sequence evolution along phylogenetic trees; see e.g. `http://bioweb2.pasteur.fr/docs/seq-gen/` or `http://tree.bio.ed.ac.uk/software/seqgen/` for details. We generated a DNA sequence alignment from a phylogenetic tree of four hypothetical taxa with equal branch lengths, using the HKY model of nucleotide substitution Hasegawa *et al.* (1985) with a uniform nucleotide equilibrium distribution, $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$, and a transition-transversion ratio of $\zeta = 2$. We generated two types of alignments. In the first alignment, the normalised branch lengths associated with the three codon positions were set to $w_i = [0.5 - \frac{c}{2}, 0.5 - \frac{c}{2}, 0.5 + c]/1.5$, where the codon offset parameter $0 \leq c \leq 0.99$ was varied in increments of 0.1. All codons had the same structure, as illustrated in Figure 3.1a. We refer to these sequence alignments as "homogeneous". The second type of alignment, which we refer to as "heterogeneous" or "subject to regional rate heterogeneity", is illustrated in Figure 3.1b. The codons have a similar structure as before. The second half of the alignment is more evolved, though, and the branch lengths are expanded by a factor of $\varsigma = 2$. In all simulations, the total length of the alignment was 1.5 Kbp.

3.4 Simulations

Our objective is to sample topology and rate state sequences $\mathbf{S}, \mathbf{R}$, their associated transition probabilities ν_S, ν_R and rate vectors ρ, the branch lengths $\mathbf{w}$ and (for the new model) the within-codon rate vector λ from the posterior distribution $P(\mathbf{S}, \mathbf{R}, \nu_S, \nu_R, \rho, \lambda, \mathbf{w}|\mathcal{D})$. To this end, we apply the Gibbs sampling scheme of (3.7–3.12), which we have described in Sections 3.2.1 and 3.2.2. Our current software has not yet been optimised for speed. Hence, to improve the convergence of the Markov chain and to focus on the aspect of interest for the present study (rate heterogeneity), we have set all states in $\mathbf{S}$ to the same tree topology without allowing for recombination: $\nu_S = 1$. We also set $K' = 2$ fixed. The model was initialised with the maximum likelihood

tree obtained with DNAML from Felsentein's PHYLIP package, available from `http://evolution.genetics.washington.edu/phylip/`. We tested the convergence of the MCMC simulations by computing the potential scale reduction factor of Gelman and Rubin Gelman and Rubin (1992) from the within and between trajectory variances of various monitoring quantities (e.g. $\mathbf{w}$, $P(R_t|\mathcal{D})$, etc.), and took a value of 1.2 as an indication of sufficient convergence.

The main objective of our study is to evaluate the performance of the proposed model that allows for within-codon rate heterogeneity; we refer to this as the "new" model. We compare its performance with a model that does not include within-codon rate heterogeneity, that is, where $\lambda = \mathbf{1}$ is constant. We refer to this as the "old" model. Note that the latter model is equivalent to the one in chapter 2 (and presented in Husmeier and Mantzaris (2008)), but with the improvement that ρ is sampled from the posterior distribution, rather than kept fixed.

In order to evaluate the performance of the methods, we want to compute the marginal posterior probability of the average effective branch length scaling for the three codon positions. The effective branch lengths are given by $\tilde{\mathbf{w}}_t = \rho_{R_t}\lambda_{I_t}\mathbf{w}_t$, where $\mathbf{w}_t$ are the normalised branch lengths. The entity of interest is

$$\Upsilon_t \;=\; \frac{||\tilde{\mathbf{w}}_t||_1}{||\mathbf{w}_t||_1} = \rho_{R_t}\lambda_{I_t} \tag{3.13}$$

which is the scaling factor by which the branch length vector $\tilde{\mathbf{w}}_t$ associated with position t deviates from the normalised branch lengths $\mathbf{w}_t$. Note that Υ_t is composed of two terms, associated with a region (ρ_{R_t}) and a codon (λ_{I_t}) effect. We are interested in the marginal posterior distribution of this factor, $P(\Upsilon|\mathcal{D}, I = k)$, for the three codon positions $I \in \{1,2,3\}$. In practice, this distribution is estimated from the MCMC sample by the appropriate marginalisation with respect to all other quantities:

$$P(\Upsilon|\mathcal{D}, I = k) \;\approx\; \frac{\sum_{i=1}^{M}\sum_{t=1}^{N}\delta_{I_t,k}\delta(\Upsilon - \rho^i_{R^i_t}\lambda^i_{I_t})}{M\sum_{t=1}^{N}\delta_{I_t,k}} \tag{3.14}$$

where the subscript t refers to positions in the alignment (of total length N), the superscript i refers to MCMC samples (sample size M), $\delta(.)$ is the delta function, the quantities on the right of its argument, $\rho^i_{R^i_t}, \lambda^i_{I_t}$, are obtained from the MCMC sample, and $\delta_{i,k}$ is the Kronecker delta. For the conventional model without explicit codon effect, we set $\lambda_{I_t} = 1/3\,\forall t$.

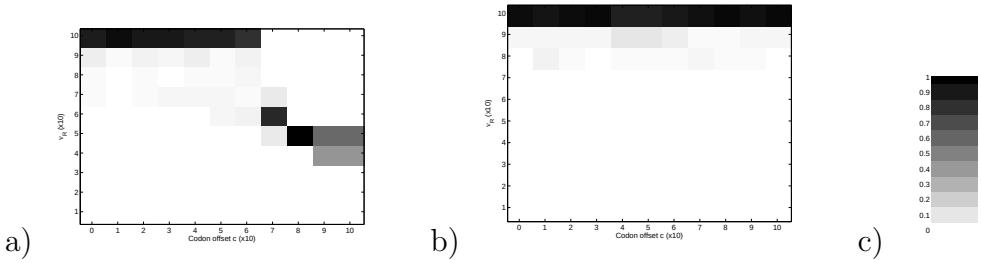


Figure 3.2: Posterior distribution of ν_R (vertical axis) for different codon offsets c (horizontal axis), where the offset indicates to what extent the nucleotide substitution rate associated with the third codon position is increased over that of the first two positions. The left panel (a) shows the results obtained with the old model, the centre panel (b) shows the results obtained with the new model. The grey levels represent probabilities, as indicated by the legend in the panel on the right (c). The distributions were obtained from a "homogeneous" DNA sequence alignment, corresponding to Figure 3.1a.

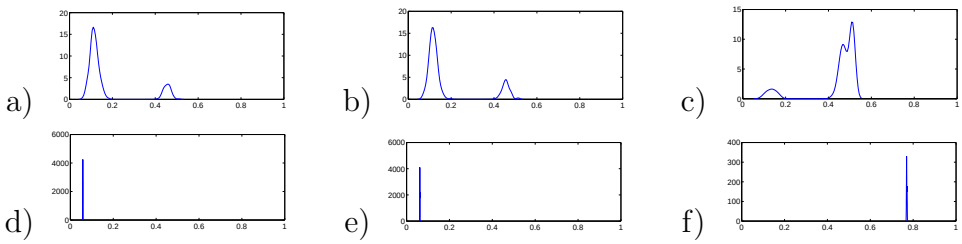


Figure 3.3: Posterior distribution (vertical axes) of the combined rate Υ_t (horizontal axes), defined in equation (3.13), for a "homogeneous" DNA sequence alignment, corresponding to Figure 3.1a, with codon offset parameter $c = 0.8$. The three columns correspond to the three codon positions. The top row shows the distribution obtained with the old model. The bottom row shows the distribution obtained with the new model. The distributions were obtained from the MCMC samples with a kernel density estimator, where the delta function in (3.14) was replaced by a Gaussian (standard deviation: a tenth of the total range).

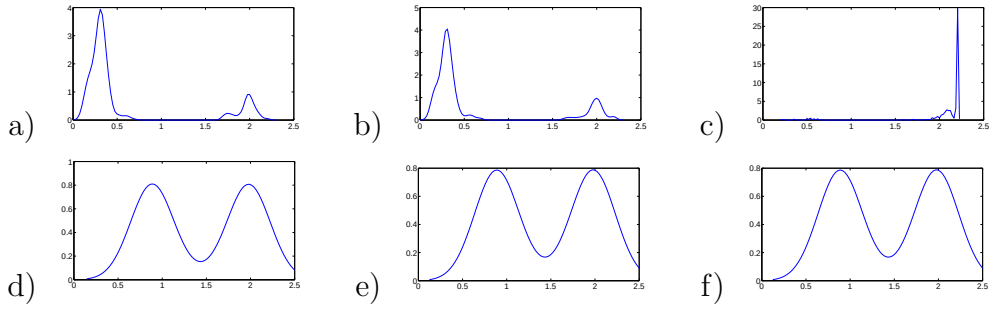


Figure 3.4: Posterior distribution (vertical axes) of the rate ρ_{R_t} (horizontal axes) for a "heterogeneous" DNA sequence alignment, corresponding to Figure 3.1b, with codon offset parameter $c = 0.8$ and regional factor $\varsigma = 2$. The three columns correspond to the three codon positions. The top row shows the distribution obtained with the old model. The bottom row shows the distribution obtained with new model. The distributions were obtained from the MCMC samples with a kernel density estimator, where the delta function in (3.15) was replaced by a Gaussian (standard deviation: a tenth of the total range).

3.5 Results

Figure 3.2 shows the posterior distribution of the (complementary) transition probability ν_R. The two models were applied to the "homogeneous" DNA sequence alignment that corresponds to the top panel in Figure 3.1. The left panel shows the results obtained with the old model, which does not explicitly include the codon effect. For small values of the offset parameter c, the posterior distribution of ν_R is concentrated on $\nu_R = 1$, which corresponds to a homogeneous sequence alignment. As the offset increases, the posterior distribution of ν_R gets shifted to smaller values, with a mode at $\nu_R = 0.5$. Note that ν_R is related to the average segment length $\bar{l}$ via the relation $\bar{l} = (1-\nu_R)\sum_l l\nu_R^{l-1} = (1-\nu_R)\frac{d}{d\nu_R}\sum_l \nu_R^l = (1-\nu_R)\frac{d}{d\nu_R}\frac{1}{1-\nu_R} = \frac{1}{1-\nu_R}$. For $\nu_R = 0.5$ we get $\bar{l} = 2$. The model has thus learned the within-codon rate heterogeneity intrinsic to the genetic code; compare with Figure 3.1. The right panel of Figure 3.2 shows the posterior distribution of ν_R obtained with the new model. Irrespective of the codon offset c, the distribution is always concentrated on $\nu_R = 1$. This correctly indicates that there is no regional rate heterogeneity in the DNA sequence alignment. Recall that the within-codon rate heterogeneity has been explicitly incorporated into the new model and, hence, need not be learned separately via ν_R and transitions between rate states R_t.

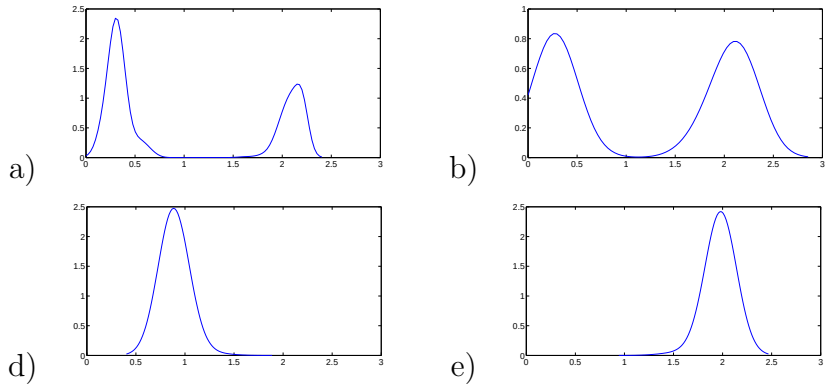


Figure 3.5: Alternative representation of the posterior distribution (vertical axes) of the rate ρ_{R_t} (horizontal axes) for the "heterogeneous" DNA sequence alignment. The figure corresponds to Figure 3.4, but shows a separation of the distributions with respect to regions rather than codon positions. The distribution of ρ_{R_t} is defined in (3.16). The two columns correspond to the two differently diverged segments in the DNA sequence alignments, with the left column representing the first 750 positions, and the right column representing the last 750 positions; the latter were evolved at double the nucleotide substitution rate. The two rows correspond to the two models. The top row shows the distribution obtained with the old model. The bottom row shows the distribution obtained with new model. The distributions were obtained from the MCMC samples with a kernel density estimator, where the delta function in (3.16) was replaced by a Gaussian (standard deviation: a tenth of the total range).

Figure 3.3 shows the posterior distribution of the scaling factor Υ_t, defined in (3.13), for the "homogeneous" DNA sequence alignment corresponding to Figure 3.1a. The columns in Figure 3.3 correspond to the three codon positions. The posterior distribution was obtained from the MCMC samples via (3.14). For the new model (bottom row of Figure 3.3), the distributions of Υ_t are unimodal and sharply peaked. This is consistent with the fact that we have no regional rate heterogeneity, and the shift in the peak locations for the third codon position clearly indicates the within-codon rate heterogeneity. For the old model (top panel of Figure 3.3), the posterior distribution is always bimodal. This is a consequence of the fact that the within-codon rate heterogeneity has to be learned via the assignment of rate states R_t to the respective codon positions. The bimodality and increased width of the distribution stem from a misassignment of rate states. Note that for an alignment of $N = 1500$ sites, 500 state transitions have to be learned to model the within-codon rate heterogeneity correctly.

Figure 3.4 is similar to Figure 3.3, but was obtained for the heterogeneous DNA sequence alignment corresponding to Figure 3.1b. For better clarity we have shown the codon site-specific posterior distributions of the rate ρ_{R_t} rather than the scale factor Υ_t, that is, in equation (3.14) we have ignored the factor $\lambda^i_{I_t}$:

$$P(\rho|\mathcal{D}, I = k) \approx \frac{\sum_{i=1}^{M}\sum_{t=1}^{N}\delta_{I_t,k}\delta(\rho - \rho^i_{R^i_t})}{M\sum_{t=1}^{N}\delta_{I_t,k}} \tag{3.15}$$

The bottom row shows the distributions obtained with the new model. They have a symmetric bimodal form. The bimodality reflects the regional rate heterogeneity. The symmetry reflects the nature of the DNA sequence alignment, which contains two differently diverged regions of equal size (see Figure 3.1b). The top panel shows the distributions obtained with the old model. The distributions are still bimodal, but the symmetry has been destroyed. This distortion results from the fact that two effects – regional and within-codon rate heterogeneity – are modelled via the same mechanism: the rate states R_t. Consequently, these two forms of rate heterogeneity are not clearly separated.

To illustrate this effect from a different perspective, Figure 3.5 shows the posterior distributions of the rate ρ_{R_t} not separated according to codon positions, but according to differently diverged regions. That is, from the MCMC sample we compute the following distribution:

$$P(\rho|\mathcal{D}, t \in r) \approx \frac{\sum_{i=1}^{M}\sum_{t=1}^{N} I(t \in r)\delta(\rho - \rho^i_{R^i_t})}{M\sum_{t=1}^{N} I(t \in r)} \tag{3.16}$$

where r represents the two regions: $r = 1$ for $1 \leq t \leq 750$, and $r = 2$ for $751 \leq t \leq 1500$, $I(t \in r)$ is the indicator function, which is one if the argument is true, and zero otherwise, and the remaining symbols are as defined below equation (3.14). The bottom panel shows the distributions obtained with the new model, where the two columns represent the two regions. The distributions are unimodal and clearly separated, which indicates that modelling regional rate heterogeneity is properly disentangled from the within-codon rate variation. The top panel shows the distributions obtained with the old model. Here, the distributions are bimodal, which results from a lack of separation between regional and within-codon rate heterogeneity, and a tangling-up of these two effects.

3.6 Discussion

We have generalised the phylogenetic FHMM of chapter 2 (also presented in Husmeier and Mantzaris (2008)) in two respects. First, by sampling the rate vector ρ from the posterior distribution with MCMC (rather than keeping it fixed) we have made the modelling of regional rate heterogeneity more flexible. Second, we explicitly model within-codon rate heterogeneity via a separate rate modification vector λ. In this way, the within-codon effect of rate heterogeneity is imposed on the model *a priori*, which should facilitate the learning of the biologically more interesting effect of regional rate heterogeneity *a posteriori*. We have carried out simulations on synthetic DNA sequence alignments, which have borne out our conjecture. The old model, which does not explicitly include the within-codon rate variation, has to model both effects with the same mechanism: the rate states R_t with associated rate factors ρ_{R_t}. As expected, it was found to fail to disentangle these two effects. On the contrary, the new model was found to clearly separate within-codon from regional rate heterogeneity, resulting in a more accurate prediction.

We emphasise that our paper describes work in progress, and we have not yet applied our method to real DNA sequence alignments. This is partly a consequence of the fact that our software has not been optimised for computational efficiency yet, resulting in long MCMC simulation runs. Note that the computational complexity of our algorithm is larger than for the model described in Lehrach and Husmeier (2009). The latter approach is based on the no-common-mechanism model of Tuffley and Steel (1997), which leads to a substantial model

simplification, though at the price of potential inconsistency problems (as discussed in chapter 2 and Husmeier and Mantzaris (2008)). The increased computational complexity of the method proposed in the present article might require the application of more sophisticated MCMC schemes, e.g. population MCMC, which will be the objective of our future work.

As a final remark, we note that a conceptually superior approach would be the modelling of substitution processes at the codon rather than nucleotide level. However, the application of this approach to standard Bayesian analysis of single phylogenetic trees has turned out to be computationally exorbitant. A generalisation to phylogenetic FHMMs for modelling DNA mosaic structures, as described in the present article, is unlikely to be computationally feasible in the near future. We therefore believe that the method we have proposed, which is based on individual nucleotide substitution processes while taking the codon structure into account, promises a better compromise between model accuracy and practical viability.

Chapter 4

Including Reversible Jump Markov Chain Monte Carlo to adapt the number of rate factors

In this chapter the methodology for inferring the number of ratefactors to be allocated along a sequence alignment with the PFHMM is developed and tested. This is done by sampling the ratefactors from the posterior distribution, using reversible jump MCMC (RJMCMC, Green (1995)). This is an improvement from chapters 3 and 2 in which a fixed number of rate states is used to fit the rate heterogeneity of the DNA sequence alignment. This restriction is removed in Lehrach and Husmeier (2009) without including the improvements made in this thesis. The chapters 2 and 3 present the improvements not found in Lehrach and Husmeier (2009). These 2 chapters made improvements by altering the model to include a branch length vector and a vector for the relative codon rate heterogeneity.

With the extra flexibility introduced to fit more features of the data, it is possible that the complexity of the model is too great for correct inference to be performed. In this chapter the model is tested on synthetic data to assess its ability to correctly infer the parameters. A failure could occur due to a lack of convergence or a vague posterior. What follows is a short introduction to the progression of the work leading to this improvement.

Husmeier (2005) introduces the FHMM for detecting recombination which this thesis makes improvements on. This is an improvement on the earlier model of Husmeier and Wright (2001) where rate heterogeneity was not taken into con-

sideration, and correct results could be gathered under conditions where the rate heterogeneity was moderate. Incorrect inference on the topology structure of the sequence alignment could be made from a large change in the levels of rate heterogeneity. Addressing this failure is the motivation behind the FHMM of Husmeier (2005) which introduces a state at every site in an independent HMM to scale the branch lengths. A short-coming of Husmeier (2005) is that it adopts the no-common-mechanism model of Tuffley and Steel (1997) allowing the branch lengths to be neglected from the inference procedure which saves on the computational expenses. The no-common-mechanism (NCM) model was then found (in chapter 2) to fail in a similar manner to parsimonious models of phylogenetics in the Felsenstein zone presented in Felsenstein (1978b). To correct for the failure, the standard model (subsection 1.2) of phylogenetics is introduced which does not allow for the branch lengths to be omitted.

The FHMM with a pre-defined size of a rate factor vector, ρ, contains the ratefactor values, ρ_i, that are allocated along the sequence alignment via the stochastic forward-backward algorithm. Here we extend the work to now include a reversible jump step for the sampling of the number of ratefactors as well as having their values sampled via MCMC. Having an RJMCMC scheme has many benefits considering that the number of ratefactors is usually not known a priori and the number of ratefactors inferred is dependent on the data. This addition along with the previous improvements avoids the pitfalls explored in chapter 2 and chapter 3 (presented in Husmeier and Mantzaris (2008) and Mantzaris and Husmeier (2009)).

4.1 Background methodology for the reversible jump MCMC scheme

Here is given a brief overview of the methodology leading to the development of the reversible jump MCMC scheme described in subsection 4.2. The background necessary for this chapter can be found in previous sections 1.3.1 (nucleotide substitution model used), 1.9 and 1.9.1 (HMM theory), 1.9.4 (FHMM theory), 1.9.5 (stochastic forward backward algorithm), 1.10.1 (Markov chain Monte Carlo), and 1.10.2 (Gibbs sampling theory). At the end of this section, the Gibbs sampling scheme with the new sampling step is presented (in eq 4.44). RJMCMC is

introduced in subsection 1.10.3.

The factorial hidden Markov model (FHMM of Husmeier (2005)) applies break points for the topologies and ratefactors along the sequence alignment. There are two a priori independent hidden Markov model chains for the topology states and the rate states. At each site in the sequence alignment, the column of nucleotides is denoted as $\mathbf{y}_t$. A topology state is denoted with S, and at each site t a topology is applied. An HMM is used to model the state sequence of topologies, and the probability of the state sequence along the alignment is given by:

$$P(\mathbf{S}) = P(S_1, \ldots, S_N) = \prod_{t=1}^{N} P(S_t|S_{t-1})P(S_1). \tag{4.1}$$

Here $\mathbf{S}$ is the vector of the topology state allocations at the sites along the alignment. The topology transition probability $P(S_t|S_{t-1})$ is dependent on the parameter ν_S, which denotes the probability of not changing state values between sites and is defined as,

$$P(S_t|S_{t-1}, \nu_S) = \nu_S^{\delta(S_t, S_{t-1})} \left(\frac{1-\nu_S}{K-1}\right)^{[1-\delta(S, S_{t-1})]}, \tag{4.2}$$

where δ represents the Kronecker delta symbol which is 1 when the topologies are equal and 0 otherwise.

The modelling of the ratestates is analogous to that of the topology states. An HMM with the hidden states representing a ratefactor value is applied to each site on the alignment. The value of each ratefactor is denoted by ρ. The vector for all of the available ratefactors that can be applied along the alignment is $\boldsymbol{\rho}$. To denote the ratestate allocation along the sequence alignment, $\mathbf{R}$, is used and t as a subscript addresses the ratefactor at that site in the alignment, R_t. The factorisation of the joint probability for the state transitions in the HMM for the rates is analogous to eq 4.1, $P(\mathbf{R}) = P(\mathbf{R}_1, \ldots, \mathbf{R}_N) = \prod_{t=1}^{N} P(\mathbf{R}_t|\mathbf{R}_{t-1})P(\mathbf{R}_1)$. The transition probability for the ratestates is given by $\nu_{\mathbf{R}}$. The transition probability definition is the same as eq 4.2 with the substitution for $\nu_{\mathbf{R}}$ made:

$$P(R_t|R_{t-1}, \nu_R) = \nu_R^{\delta(R_t, R_{t-1})} \left(\frac{1-\nu_R}{\tilde{K}-1}\right)^{[1-\delta(R_t, R_{t-1})]}. \tag{4.3}$$

The new symbol included here is $\tilde{K}$ and denotes the number of ratefactors in $\boldsymbol{\rho}$. Each site in the alignment, t, has a ratefactor applied to it for scaling the branch lengths, $R_t \in \rho_1, \ldots, \rho_{\tilde{K}}$. Both of the FHMM state transition parameters ν_S and ν_R have a beta distribution as a prior that is set to be non-informative

(uniform distribution). The prior distribution is discussed in detail in Husmeier and McGuire (2003) and in section 1.9.3.

The vector of branch lengths is denoted by $\mathbf{w}$ and the emission probabilities for the sites of the sequence alignment, $\mathbf{y}_t$, are

$$P(\mathbf{y}_t|S_t,\mathbf{w}). \tag{4.4}$$

How this is computed is described in section 1.5. An unidentifiability problem between the ratefactors and the branch lengths can occur and the solution applied in chapter 2 is used here. The vector of branch lengths is sampled from a Dirichlet distribution producing normalised vectors, $|\dot{\mathbf{w}}| = 1$, where all the lengths are not negative. With the normalisation constraint on the branch lengths the ratefactor, ρ, is the only parameter taking on the role of scaling the expected number of mutations. The new vector of branch lengths, $\mathbf{w}_k^*$ proposed from a Dirichlet distribution is conditional on the current vector,

$$Q(\mathbf{w}_k^*|\mathbf{w}_k) \propto \prod_i [w_{ki}^*]^{\alpha w_{ki}-1} \ . \tag{4.5}$$

The variable α controls the variance of the distribution conditional on the original branch length vector. This is tuned in the burnin phase of the simulation to achieve the desired average acceptance percentage of values between 30 and 70 percent. These percentages are found in many of the referenced papers.

The codon rate heterogeneity vector λ is a vector with 3 values for each position of the amino acid encoding. Each position corresponds to a ratefactor in the codon triplet. As with the branch lengths the relative rates for the codons are expressed as a normalised vector, $|\lambda| = 1$, resulting in identifiable solutions. A site t corresponds to one of the three positions in the vector, $I_t = i \in 1,2,3$. The proposal mechanism for the values of this vector is analogous to that for the branch lengths shown in eq 4.5. The emission probability of the model discussed in chapter 3 which includes λ is,

$$P(\mathbf{y}_t|S_t,\rho\mathbf{w}\lambda). \tag{4.6}$$

From the study done in chapter 3 and presented in Mantzaris and Husmeier (2009), it is seen that the exponential prior excessively penalises higher ratefactor values. The higher values for the ratefactors are needed to compensate the normalised vector of codon lengths. The solution replaces the exponential prior

with a uniform prior. The prior on the ratefactors will be discussed later in this chapter when the model developed here is introduced.

Inference with the HMM delivers the marginal posterior distribution of each topology along sites of the alignment (equivalently for the ratefactors allocated),

$$P(S_t|\mathcal{D}) = \sum_{S_t}\cdots\sum_{S_{t-1}}\sum_{S_{t+1}}\cdots\sum_{S_N}P(\mathbf{S}|\mathcal{D}) \quad (4.7)$$

$$P(R_t|\mathcal{D}) = \sum_{\mathbf{R}_1}\cdots\sum_{\mathbf{R}_{t-1}}\sum_{\mathbf{R}_{t+1}}\cdots\sum_{\mathbf{R}_N}P(\mathbf{R}|\mathcal{D}) \quad (4.8)$$

The distribution for the sequence of topologies, $\mathbf{S}$, and for the sequence of rate states is obtained by marginalising over the remaining parameters:

$$P(\mathbf{S}|\mathcal{D}) = \sum_{\mathbf{R}}\int P(\mathbf{S},\mathbf{R},\nu_S,\nu_R,\mathbf{w},\lambda_{1,2,3}|\mathcal{D})d\nu_S d\nu_R d\mathbf{w} d\lambda_{1,2,3} \quad (4.9)$$

$$P(\mathbf{R}|\mathcal{D}) = \sum_{\mathbf{S}}\int P(\mathbf{S},\mathbf{S},\nu_S,\nu_R,\mathbf{w},\lambda_{1,2,3}|\mathcal{D})d\nu_S d\nu_R d\mathbf{w} d\lambda_{1,2,3}. \quad (4.10)$$

The posterior is proportional to the likelihood:

$$P(\mathbf{S},\mathbf{R},\nu_S,\nu_R,\mathbf{w},\lambda|\mathcal{D}) \propto P(\mathcal{D},\mathbf{S},\mathbf{R},\nu_S,\nu_R,\mathbf{w},\lambda) \quad (4.11)$$

and is equal to,

$$\begin{aligned} &P(S_1)P(R_1)P(\nu_S)P(\nu_R)P(\lambda)\times \\ &\prod_{t=1}^{N}P(\mathbf{y}_t|S_t,R_t\mathbf{w}\lambda)\prod_{t=2}^{N}P(S_t|S_{t-1},\nu_S)\prod_{t=2}^{N}P(R_t|R_{t-1},\nu_R). \end{aligned} \quad (4.12)$$

The state transition terms for the rate state and topology states are defined in eq 4.1, eq 4.2 and eq 4.3.

These summations and integrations of eq 4.9 are not analytically tractable. The summations over the state space of the rates and topologies are done within the HMM using the stochastic forward backward algorithm described in section 1.9.5. The parameters are sampled from the posterior distribution using MCMC. The Gibbs sampling scheme described in subsection 1.10.2 is,

$$\mathbf{S}^{(i+1)} \sim P(\cdot|\mathbf{R}^{(i)},\nu_S^{(i)},\nu_R^{(1)},\mathbf{w}^{(i)},\lambda^{(i)},\rho^{(i)},\mathcal{D}) \quad (4.13)$$

$$\mathbf{R}^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)},\nu_S^{(i)},\nu_R^{(1)},\mathbf{w}^{(i)},\lambda^{(i)},\rho^{(i)},\mathcal{D}) \quad (4.14)$$

$$\nu_S^{(i+1)} \sim P(\cdot|\mathbf{R}^{(i+1)},\mathbf{S}^{(i+1)},\nu_R^{(1)},\mathbf{w}^{(i)},\lambda^{(i)},\rho^{(i)},\mathcal{D}) \quad (4.15)$$

$$\nu_R^{(i+1)} \sim P(\cdot|\mathbf{R}^{(i+1)},\mathbf{S}^{(i)},\nu_S^{(i+1)},\mathbf{w}^{(i)},\lambda^{(i)},\rho^{(i)},\mathcal{D}) \quad (4.16)$$

$$(\mathbf{w}^{(i+1)},\lambda^{(i+1)}) \sim P(\cdot|\mathbf{R}^{(i+1)},\mathbf{S}^{(i+1)},\nu_S^{(i+1)},\nu_R^{(i+1)},\rho^{(i)},\mathcal{D}) \quad (4.17)$$

$$\rho^{(i+1)} \sim P(\cdot|\mathbf{R}^{(i+1)},\mathbf{S}^{(i)},\nu_S^{(i+1)},\mathbf{w}^{(i+1)},\lambda^{(i+1)},\mathcal{D}). \quad (4.18)$$

The variable, i, denotes the iteration number in which the simulation is running in. The first equation and the second one are sampled via the stochastic forward-backward algorithm. The 3rd and 4th are sampled as in eq 1.79 from the beta distribution. The vectors $\mathbf{w}$ and λ are sampled using Metropolis-Hastings. In the last equation shown, eq 4.44, the ratefactors are sampled as in the previous chapter 3 with Metropolis-Hastings, and the model is improved by utilising RJMCMC. Table 4.1 shows the symbols used in the rest of this chapter and is useful as a quick reference.

4.2 Application of the RJMCMC scheme

The introduction to reversible jump Markov chain Monte Carlo (RJMCMC) is given in subsection 1.10.3. The model described here changes the use of the prior. The choice of the prior probability on the ratefactors in eq 4.27 is different to the prior used in Lehrach (2008) where the exponential prior on the values is used. The exponential prior (informative prior) is not used here because of the strong penalisation it introduces for moderately large ratefactor values. The exponential prior was suitable when the model was not taking into account the codon rate variation, modelled with a normalised rate vector for the positions, λ. The values of the ratefactors are required to be larger to overcome the effect of multiplying them with the values of the vector of the codon rate heterogeneity. The uniform prior (non-informative prior) is chosen for the ratefactors. The bounds for the ratefactors used here are on the log scale of -3 to 2 (since these provide an adequate range for the ratefactor values), and distribution of the prior on the ratefactors is;

$$u = \log\rho \tag{4.19}$$

$$P(\mathbf{u}) = P(\log\rho) \tag{4.20}$$

$$P(\mathbf{u}) = \frac{1}{\Delta^{\tilde{K}}}\Pi_{k=1}^{\tilde{K}}\mathbb{I}(\log\rho_{\min} \leq u_k \leq \log\rho_{\max}) \tag{4.21}$$

$$\log\rho_{\min} = -3, \log\rho_{\max} = 2 \tag{4.22}$$

$$\Delta = \rho_{\max} - \rho_{\min}. \tag{4.23}$$

A ratefactor value of zero is given for values outside the boundary and which is produce by using the indicator function $\mathbb{I}$. Lehrach (2008) explored the use of three different priors: the uniform distribution, a normal distribution and the

symbols used	description
$\mathcal{D}$	The DNA sequence alignment which is the data and considered as an array of nucleotides from $1 \ldots N$ and each column is indexed by t.
$\mathbf{y}_t$	The vector of nucleotides at site t in the alignment.
$\mathbf{S}$	Vector of the selected hidden topology states at each site in the alignment, sampled in the Gibbs simulation.
$\mathbf{R}$	Vector of rate states allocated to each site in the alignment. Each rate state is an index to the rate factor vector ρ.
$\mathbf{w}$	The vector of branch lengths as a normalised vector for each length in the phylogenetic tree.
ρ	The vector of ratefactor values.
ρ_i	An individual ratefactor value, that scales uniformly all the branch lengths.
λ	Normalised vector of relative rates between the 3 codon positions.
ν_S	The transition probability for a topology state to not change between the sites of the sequence alignment.
ν_R	The transition probability for a rate state value to not change between the sites of the sequence alignment.

Figure 4.1: The symbols used in the reversible jump MCMC scheme

The symbols listed in the first column will be used in the RJMCMC sampling scheme. A brief description of the quantities is given in the second column.

even-numbered order statistic. All three of these priors showed similar results and the simplest one (being the uniform distribution) is chosen here.

The ratefactors proposed are between the intervals shown in eq 4.22. The boundaries for this interval, on the log scale, are chosen so that the range of reasonable ratefactors are only available and undefined values as well as redundant values are excluded. Undefined ratefactors are those with negative values and redundant values are those beyond the value close enough to the stationary distribution for the nucleotide distribution.

The distribution on the ratefactors is uniform on the log scale, $P(\log\rho)$ and is defined in eq 4.19-4.23. A birth move will take this form:

$$\left(\frac{P(\mathcal{D}|\theta,\rho_{\tilde{K}'})}{P(\mathcal{D}|\theta,\rho_{\tilde{K}})}\times\frac{P(\rho_{\tilde{K}'})}{P(\rho_{\tilde{K}})}\times\frac{P(death)P(\tilde{K}'\rightarrow\tilde{K})}{P(birth)Q(\rho')P(\tilde{K}\rightarrow\tilde{K}')}\times\text{Jacobian}\right). \tag{4.24}$$

Here we have used θ to represent all the model parameters except that of the rates to focus on the change of the number of rate factors (from 1 to 2 ratefactors in the ratefactor vector; $\tilde{K}'=2,\tilde{K}=1$). The death move is the inverted case of the birth move as each of the fractional components have the nominator and denominator swapped.

The acceptance probability for the RJMCMC scheme is as follows. The terms for the LR, PR and IPPR are shown and the Jacobian value is omitted as it takes on the value of 1 in every case (and is explained after eq 4.30). The ratefactor state sequence $\mathbf{R}'$ denotes the new sampled state allocations of the ratefactors along the FHMM, which is done via the stochastic forward backward algorithm which is presented in subsection 1.9.5. The state allocations for the topology, $\mathbf{S}$, at a site are not changed in this stage of the Gibbs sampling scheme. Variables with an apostrophe are the proposed new values.

$$\begin{aligned} LR &= \frac{P(\mathbf{R}',\mathcal{D}|\theta,\tilde{K}+1,\mathbf{S},\rho')}{P(\mathbf{R},\mathcal{D}|\theta,\tilde{K},\mathbf{S},\rho)} \\ &= \frac{P(\mathbf{R}'|\mathbf{S},\mathcal{D},\tilde{K}+1,\theta,\rho')}{P(\mathbf{R}|\mathcal{D},\tilde{K},\theta,\mathbf{S},\rho)}\times\frac{P(\mathcal{D}|\mathbf{S},\tilde{K}+1,\theta,\rho')}{P(\mathcal{D}|\mathbf{S},\tilde{K},\theta,\rho)} \end{aligned} \tag{4.25}$$

The terms are separated in this way so that they can cancel out with the analogous term in the IPPR of eq 4.30; the posterior distribution of the ratestate sequence. For the prior, we have the probability on the number of components and the probability for the given ratefactor vector. The probability on the number of

ratefactors is given by the Poisson distribution, $Poiss(\tilde{K})$. The distribution on the ratefactor vector, ρ, is defined in equations 4.19- 4.23. The substitution for the number of ratefactors on the uniform scale is,

$$P(\tilde{K}, \log_{\rho}) = P(\tilde{K})P(\log\rho|\tilde{K}) = P(\tilde{K})P(\log\rho)^{\tilde{K}} = Poiss(\tilde{K})\left(\frac{1}{\Delta}\right)^{\tilde{K}}. \tag{4.26}$$

The prior ratio of a birth step of $\tilde{K}$ to $\tilde{K}+1$ number of components in the ratefactor vector ρ is,

$$\begin{aligned} PR &= \frac{Poiss(\tilde{K}+1)P(\log\rho)^{\tilde{K}+1}}{Poiss(\tilde{K})P(\log\rho)^{\tilde{K}}} \\ &= \frac{Poiss(\tilde{K}+1)P(\log\rho)}{Poiss(\tilde{K})} \\ &= \frac{Poiss(\tilde{K}+1)\frac{1}{\Delta}}{Poiss(\tilde{K})}. \end{aligned} \tag{4.27}$$

The proposal distribution is made in such a way as to cancel out with the prior. The canceling of terms is motivated by simplicity and for there to be a higher acceptance ratio in the simulation to assist convergence.

Assume there are $\tilde{K}$ rate states, i.e. $\dim(\rho) = \tilde{K}$. A birth move consists of the following steps. First, we select a birth move with probability $b_{\tilde{K}}$. Next, sampling a new ratefactor from the proposal distribution, which is selected to be the same as the prior distribution $P(\log\rho)$, defined in eq 4.19- 4.23. There are $(\tilde{K}+1)!$ ways of assigning $\tilde{K}+1$ rate factors to $\tilde{K}+1$ states. These assignments of ratefactors are all equivalent in that both the likelihood and the prior distribution are invariant with respect to label switching (described in subsection 4.2.1). One particular assignment with probability $1/(\tilde{K}+1)!$ can be chosen. Finally, a sample for new rate states from the posterior distribution $P(\mathbf{R}|\mathcal{D}, \tilde{K}, \theta, \mathbf{S}, \rho)$ with the dynamic programming scheme is done as described in subsection 1.9.5. The overall probability of a birth move from $\tilde{K}$ components to $\tilde{K}+1$ components is,

$$Q_{birth} = \frac{b_{\tilde{K}}P(\log\rho)}{(\tilde{K}+1)!}P(\mathbf{R}'|\mathcal{D}, \tilde{K}+1, \theta, \mathbf{S}, \rho'). \tag{4.28}$$

Now considering the complementary death move. Given that there are $(\tilde{K}+1)$ states, a death move is selected with probability $d_{\tilde{K}+1}$. Next, randomly selecting one of the $(\tilde{K}+1)$ components to be killed, with probability $1/(\tilde{K}+1)$. In the resulting configuration with a $\tilde{K}$ number of states, there are $\tilde{K}!$ possibilities of assigning the $\tilde{K}$ rate factors to the $\tilde{K}$ states. The invariance with respect to label

switching (described in subsection 4.2.1) does not affect the likelihood or the prior distribution and a random assignment with probability $1/\tilde{K}!$ is chosen. New rate state values are sampled from the posterior distribution $P(\mathbf{R}|\mathcal{D},\tilde{K},\theta,\mathbf{S},\rho)$ using the stochastic forward backward algorithm described in section 1.9.5. The overall probability of the complementary death move from $(\tilde{K}+1)$ components to $\tilde{K}$ components is,

$$Q_{death} = \frac{d_{\tilde{K}+1}}{(\tilde{K}+1)\tilde{K}!}P(\mathbf{R}|\mathcal{D},\tilde{K},\theta,\mathbf{S},\rho) = \frac{d_{\tilde{K}+1}}{(\tilde{K}+1)!}P(\mathbf{R}|\mathcal{D},\tilde{K},\theta,\mathbf{S},\rho). \quad (4.29)$$

Combining equations (4.28) and (4.29), for the inverse proposal probability ratio *IPPR*:

$$\begin{aligned} IPPR &= \frac{Q_{death}}{Q_{birth}} = \frac{d_{\tilde{K}+1}(\tilde{K}+1)!}{b_{\tilde{K}}P(\log\rho)(\tilde{K}+1)!}\frac{P(\mathbf{R}|\mathcal{D},\tilde{K},\theta,\mathbf{S},\rho)}{P(\mathbf{R}'|\mathcal{D},\tilde{K}+1,\theta,\mathbf{S},\rho')} \\ &= \frac{d_{\tilde{K}+1}}{b_{\tilde{K}}P(\log\rho)}\frac{P(\mathbf{R}|\mathcal{D},\tilde{K},\theta,\mathbf{S},\rho)}{P(\mathbf{R}'|\mathcal{D},\tilde{K}+1,\theta,\mathbf{S},\rho')}. \end{aligned} \quad (4.30)$$

The final factor to be derived is the Jacobian. For the birth and death moves performed, the absolute value of Jacobian's determinant is taken, and it equals 1 at all times (the absolute value ensures that negative numbers are rejected). Merge and split moves would result in the Jacobian taking on other values, but these are not performed. The work of Boys and Henderson (2002) shows that in their RJMCMC scheme applied to DNA, the birth and death moves are accepted more often than merge and split moves are. They are conceptually more simple, require less computational effort, show better mixing, and work well in the application to DNA (single sequences) where the application is similar. In the birth and death moves the new components proposed are independent of all the other values within the current vector, and in the death moves as well. The terms in the Jacobian for a ratefactor has a zero valued partial derivative when not differentiated with itself and a value of 1 when differentiated with itself. The non-diagonal elements of the matrix are 0 by definition of their independence. The diagonal values of the Jacobian are all 1, since the partial differential with respect to itself equals 1, and the determinant of the identity matrix that results equals 1. The elements in the matrix are,

$$\frac{\partial\rho_i}{\partial\rho_j} = \mathbf{l}\delta_{ij} \quad (4.31)$$

where for $i \neq j$ the entry is 0 since the delta is the Kronecker delta and where for $i = j$ the entry is $l = 1$. The Jacobian in every case is the identity matrix whose

determinant is 1,

$$\det \mathbf{J} = 1. \tag{4.32}$$

The appendix in section A.16 and section A.12 demonstrate in more detail these concepts.

The RJMCMC algorithm can perform 3 different operations; the birth move $b_{\tilde{K}}$, death move $d_{\tilde{K}+1}$, and the relocation step $m_{\tilde{K}}$. The probabilities for choosing between these operations is given by the following 3 equations,

$$b_{\tilde{K}} = c \times min\{1, Poiss(\tilde{K}+1)/Poiss(\tilde{K})\} \tag{4.33}$$

$$d_{\tilde{K}} = c \times min\{1, Poiss(\tilde{K}-1)/Poiss(\tilde{K})\} \tag{4.34}$$

$$m_{\tilde{K}} = 1 - b_{\tilde{K}} - d_{\tilde{K}}. \tag{4.35}$$

These equations apply a probability to the three possible moves of the RJMCMC sampler scheme. The value of c is chosen to be 0.4 following the choice made in Suchard *et al.* (2003). With these three probabilities that add to 1, a random selection proportional to these probabilities is made for which step is chosen. In this way, the ratio $\frac{d_{K+1}}{b_K} = \frac{P(K)}{P(K+1)}$ cancels out against the prior probability ratio.

Combining equations 4.25(for the likelihood), 4.27 (for the prior), 4.30 (for the Hastings factor), and 4.32(for the Jacobian), we have the product of terms,

$$LR \times PR \times IPPR \times |\mathbf{J}| = \frac{P(\mathcal{D}|\mathbf{S}, \tilde{K}+1, \theta, \rho')}{P(\mathcal{D}|\mathbf{S}, \tilde{K}, \theta, \rho)}. \tag{4.36}$$

This convenient form left for the acceptance probability is,

$$A_b = Min\{1, \frac{P(\mathcal{D}|\mathbf{S}, \tilde{K}+1, \theta, \rho')}{P(\mathcal{D}|\mathbf{S}, \tilde{K}, \theta, \rho)}\}. \tag{4.37}$$

The derivation for a death move is analogous to the one above. Given $\tilde{K}$ components and a death move is selected with probability $d_{\tilde{K}}$, and a birth move is selected with probability $b_{\tilde{K}-1}$ a similar derivation is made. The death LR ratio is:

$$LR = \frac{P(\mathbf{R}'|\mathbf{S}, \mathcal{D}, \tilde{K}-1, \theta, \rho')}{P(\mathbf{R}|\mathcal{D}, \tilde{K}, \theta, \mathbf{S}, \rho)} \times \frac{P(\mathcal{D}|\mathbf{S}, \tilde{K}-1, \theta, \rho')}{P(\mathcal{D}|\mathbf{S}, \tilde{K}, \theta, \rho)} \tag{4.38}$$

and the PR ratio is:

$$PR = \frac{Poiss(\tilde{K}-1)}{Poiss(\tilde{K})P(\log \rho)}. \tag{4.39}$$

The overall probability of a birth move from $\tilde{K}-1$ components to $\tilde{K}$ components is:

$$Q_{birth} = \frac{b_{\tilde{K}-1}P(\log\rho)}{(\tilde{K})!}P(\mathbf{R}|\mathcal{D},\tilde{K},\theta,\mathbf{S},\rho), \tag{4.40}$$

and the probability of the death move of $\tilde{K}$ to $\tilde{K}-1$ is,

$$Q_{death} = \frac{d_{\tilde{K}}}{(\tilde{K})!}P(\mathbf{R}'|\mathcal{D},\tilde{K}-1,\theta,\mathbf{S},\rho'). \tag{4.41}$$

Lastly for the IPPR of the death move,

$$IPPR = \frac{Q_{birth}}{Q_{death}} = \frac{b_{\tilde{K}-1}P(\log\rho)}{d_{\tilde{K}}}\frac{P(\mathbf{R}|\mathcal{D},\tilde{K},\theta,\mathbf{S},\rho)}{P(\mathbf{R}'|\mathcal{D},\tilde{K}-1,\theta,\mathbf{S},\rho')}, \tag{4.42}$$

and finally the equation for the acceptance of a death move from $\tilde{K}$ to $\tilde{K}-1$ is,

$$A_d = Min\{1,\frac{P(\mathcal{D}|\mathbf{S},\tilde{K}-1,\theta,\rho')}{P(\mathcal{D}|\mathbf{S},\tilde{K},\theta,\rho)}\}. \tag{4.43}$$

The relocation step performs an MCMC simulation for the rate state values in the Gibbs sequence in eq 4.44. The pseudocode in the appendix A.15 shows the steps for the MCMC simulation of the ratefactor vector. Since the dimensionality of the ratefactor vector varies during the Gibbs simulation, the number of ratefactors is indicated in the subscript of $\rho_{\tilde{K}}$. The equation is,

$$\rho^{(i+1)}_{\tilde{K}^{(i+1)}} = P(\cdot|\mathbf{R}^{(i)},\mathbf{S}^{(i)},\nu_S^{(i+1)},\mathbf{w}^{(i+1)},\lambda^{(i+1)},\mathcal{D}). \tag{4.44}$$

4.2.1 Background on label switching in the RJMCMC sampler

The ratefactor vector, ρ, with its $\tilde{K}$ number of ratefactors, can have $\tilde{K}!$ permutations of these values. Each permutation has the same value in the likelihood function and prior. These are multiple modes based on the permutations of the elements in the ratefactor vector. This is because the prior and likelihood are invariant towards the switching of the labels of the ratefactor vector components.

Considering a scenario where there are two ratefactors in the ratefactor vector with values of 1 and 2; $\rho = [1,2]$. If subsequently the rate factor of value 2 is removed in a death move, and then later on in the simulation the same value is proposed in a birth move and accepted to be placed as the first ratefactor element. The result is then $\rho = [2,1]$. The ratefactor vector has effectively the same contribution as before in the model but the labelling of the components has changed.

The identifiability problem due to label switching in models where the prior and likelihood functions are invariant towards the permutations of the labels is explored in Jasra *et al.* (2005), Green (1995), Lehrach (2008), and chapter 6 of Marin and Robert (2007). One approach to address this issue is by introducing an artificial identifiability constraint (AIC) on the parameters of the model, restricting the sampler to a single mode:

$$P(\rho|\tilde{K}) = \mathbb{I}(\rho_1 < \rho_2 ... < \rho_N)(\tilde{K}!)\Pi_{i=1}^{N} Q(\rho_i). \tag{4.45}$$

Here the indicator function is used, $\mathbb{I}(\cdot)$, which equals 1 when the arguments are true and 0 otherwise.

This is a naive approach because there are consequences on the inference imposed by the AIC. Exploration of better posterior configurations can be prevented as the sampling procedure is more difficult with the AIC. Where it would be desirable to have a single mode which the sampler is restricted to, the constraints may include parts of other modes and the high density posterior regions may occur at boundaries between parameters. The AIC may also hinder the information contained in the prior. In distributions of more than one variable the distance between parameters in corresponding distributions may not be the same for each variable. For instance, the means of two normal distributions may be close but have variances that differ greatly in value. Chapter 6 of Marin and Robert (2007) discusses these topics in more depth.

An alternative approach to the AIC is that the constraint is applied after the simulations have finished, and the relabelling is done on the sampled parameters which were not restricted. The posterior reordering can be done by selecting the *maximum a posteriori* (MAP) as a reference point (a pivot where most samples will be taken from the region around this mode). This reordering and switching to one mode is not ideal, and is discussed in Jasra *et al.* (2005) and Celeux *et al.* (2000).

The proposed solution is to take the values from the sampling scheme as being label independent. By having the labels not influence conclusions from the sampled values the inference procedure is invariant towards label switching. Here the ratefactors considered are independent of labels as only the actual values associated to the components are considered. At each site in the alignment the values allocated to that site over the duration of the simulation is averaged over and this average value is used. Figure 4.2 depicts this approach taken, and eq 4.48

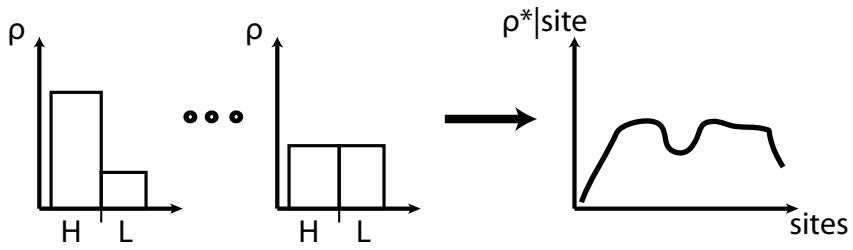


Figure 4.2: Ratefactor relabelling

The figure shows the effect of relabelling from the RJMCMC sampler. The left two hypothetical charts show the value of the ratefactor vector positions that can have high (H) and low (L) value labels. The three circles between them show a later stage in the Gibbs sampling scheme where a death move has occurred from there being two ratefactors in the ratefactor vector to only one ratefactor to one after a death move. A death move forces the remaining ratefactor to take on a value to satisfy both the high and low value regions resulting in an 'average' value over all the positions. The arrow shows a pictorial result from the completed simulation where the ratefactor values along the sites in the alignment are presented. Since only the values of the allocated ratefactors are taken for each position during the simulation, the labellings do not affect the inferred ratefactor value at each position. The RJMCMC scheme would only cause a problem with labelling if a certain ratefactor with a label were associated with a set of the sites in the data. Eq 4.48 displays the equation used to achieve this.

defines how this is computed. From the values of the ratefactors, the mean, the standard deviation and the percentiles are computed. These do not depend on the labels, and the approach is therefore invariant with respect to the label switching.

4.3 Data

The purpose of the synthetic data study is to test that the RJMCMC inference scheme is working properly. A range of problems in terms of difficulty are examined. The MATLAB programs used in Mantzaris and Husmeier (2009) were extended to incorporate the transdimensional sampling of the ratefactor values.

In the work of Mantzaris and Husmeier (2009) (chapter 3) and Husmeier and Mantzaris (2008) (chapter 2) synthetic sequence alignments were generated using

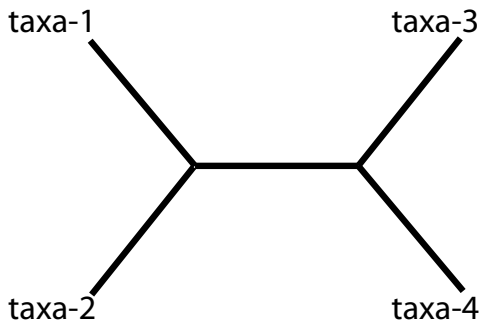


Figure 4.3: Codon Model

A phylogenetic tree of 4 taxa with taxa 1 and 2 being adjacent. Here the branch lengths are all of equal lengths.

the program SEQGEN, Rambaut (1996), available online as a web service:

$$http://bioweb2.pasteur.fr/docs/seq-gen/, \tag{4.46}$$

and the stand alone application supplied by the author,

$$http://tree.bio.ed.ac.uk/software/seqgen/. \tag{4.47}$$

The alignments generated for this study were created using SEQGEN and also from MATLAB programs the authors wrote for generating sequence alignments according to a phylogenetic tree. Both programs were set to use the HKY model of nucleotide substitution (Hasegawa *et al.* (1985)) set with the transition/transversion value of 2 as presented in subsection A.3. The equilibrium distribution for the nucleotides was set to uniform $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$ which results in a model identical to the Kimura model described in subsection 1.3.3. Fig 4.3 shows a phylogenetic tree from 4 taxa, with taxa 1 and 2 being adjacent.

4.4 Synthetic Sequence Alignments

The various synthetic alignments are produced for testing the operation of the new model's capability at inferring the rate parameters and topologies along the alignment. The relative rates of mutation at the codon level is uniform across the alignment $\lambda_{1,2,3} = \frac{1}{3}$, and the branch lengths are kept uniform for the 5 values $\mathbf{w}_{1,\ldots,5} = \frac{1}{5}$. The synthetic alignments produced are varied to test the ability of the model to fit topology changes along the alignment as well as changes in the rate of mutation.

4.4.1 Alignment 1

A sequence alignment of 1.5Kbp is generated with 3 sections. Each section is a continuous strand of 500bp each having the same topology as Fig 4.3 where taxa 1 and 2 are adjacent. In the first 500bp, $\mathbf{y}_{1...500}$ the rate of mutation is 1 which creates a branch length vector of $\mathbf{w}_{1,...,5} = \frac{1}{5}$. The columns $\mathbf{y}_{500...1000}$ are scaled by 2 with the ratefactor creating a branch length vector of $\mathbf{w}_{1,...,5} = \frac{2}{5}$. The last base pairs of the alignment $\mathbf{y}_{500...1500}$ are scaled by 0.5 (the ratefactor value) creating a branch length vector of $\mathbf{w}_{1,...,5} = \frac{1}{10}$.

4.4.2 Alignment 2

This synthetic DNA sequence alignment generated is similar to the alignment in the previous subsection 4.4.1. The length of the alignment is the same and the structure of rate heterogeneity is the same as well. The difference is that a recombination event is included (change in the topology along the alignment). The topology break point is at site 500 of the alignment. The topology bringing taxa 1 and taxa 3 to be adjacent is used for generating the first 500bp and the rest of the alignment with the topology grouping taxa 1 and 2 to be adjacent. The branch length vectors for each topology is still uniform along all the lengths. With the ratefactors the lengths are scaled to values $\mathbf{w}_{1,...,5} = \frac{1}{5}$, $\mathbf{w}_{1,...,5} = \frac{2}{5}$, and $\mathbf{w}_{1,...,5} = \frac{1}{10}$ in the 3 respective regions. The relative codon rate vector $\boldsymbol{\lambda}$ is also uniform along the whole alignment $\boldsymbol{\lambda}_{1...3} = \frac{1}{3}$ as was for the Alignment 1 in the previous subsection 4.4.1. This alignment is used to further reinforce that the model is capable of learning the right number of rate states.

4.4.3 Alignment 3

The previous two synthetic sequence alignments (subsection 4.4.1 and subsection 4.4.2) were used to test the model's ability sample the number of rate states which reflects the data generating process of the data. Alignment 2 in subsection 4.4.2 tests the model's ability to sample the correct number of rate states, as well as allow a change in the topology along the alignment. The topology change happens at site 500, where a ratefactor state change also occurs. This alignment number 3 is a variation to test the model's ability to fit to data where the rate state and topology state change are not occurring at the same site on the alignment. The only difference here is that the topology structure of the alignment

is the grouping of taxa 1 and 3 for the sites $1 \ldots 750$, and the topology grouping taxa 1 and 2 for the sites $751 \ldots 1500$. The ratefactors along the alignment have break points at sites 500, 1000, and 1500 with rate factors values 1, 2, and 0.5 respectively.

4.5 Simulations

A histogram of the discrete number of ratefactors sampled from the RJMCMC scheme, $\tilde{K}_i$ is made. The marginal posterior distributions for the topology at each site, $P(S_t|\mathcal{D})$, is plotted along the N sites of the alignment for each of the 3 possible topologies. The posterior probabilities of the HMM transition parameters $\nu_{R,S}$ are plotted along the number of Gibbs iterations as well as the log likelihood. The individual ratefactors, ρ_i where $i \in 1 \ldots \tilde{K}$, are not plotted individually. The mean of the sampled ratefactor values for each Gibbs iteration is plotted as the mean of the posterior probabilities at each site on the alignment with the 50 and 95 percent credibility intervals plotted as well.

$$P(\rho_t|\mathcal{D}) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{\tilde{K}} \rho_j \delta(R_t^i, j)}{N} \tag{4.48}$$

is the equation used for finding the ratefactor value at each site where δ is the Kronecker delta, and N is the number of Gibbs iterations. The Bayesian credibility intervals are obtained by measuring the standard deviation in the samples of the ratefactors at each site in the alignment. This is done using the standard formula:

$$\bar{x} \pm \frac{c \times \sigma_x}{\sqrt{n}}. \tag{4.49}$$

Here $x \sim P(\rho_t|\mathcal{D})$, n is the number of samples and c takes the values [0.6745,1.96] for the 50 and 95 percent credibility intervals respectively.

To monitor the convergence of the sampling procedure, Gelman and Rubin (1992), the potential scale reduction factors (PSRF) are measured for the parameters after the burnin phase which had values less than 1.2. The quantities measured are from the trajectory of samples of the branch lengths for each topology as well as the relative rates for the codon structure, the number of rate states, and the values of the rate states. The simulations are all done with random initial configurations.

4.6 Results

In this section the results of the improved model used on synthetic DNA sequence alignments are presented. The model performs correct inference with short sequences, topology changes (recombination events), and ratestate changes.

4.6.1 Alignment 1

Subsection4.4.1 describes the alignment examined under the extended FHMM with trans dimensional sampling for the number of ratefactors. Fig 4.4 summarises the results. In subfigure a) the posterior probabilities of the topologies are shown along the sites in the alignment. The first subplot represents the topology where the taxa labelled 1 and 2 are adjacent to each other which is the topology used to create the data sequence alignment. The other topologies contribute a negligible posterior probability. The subfigure b) shows the mean ratefactor values during the simulation along the sites of the sequence alignment with the 50 and 95 percent credibility intervals plotted (eq 4.48). The two break points separating the three regions of different values of rate heterogeneity are clearly distinct. The credibility intervals do not diverge from the mean at the sites around the break point either. The correct values of 1, 2, and 0.5 are clearly seen. Subfigure c) shows the histogram of the number of ratefactors, $\tilde{K}$ that the RJMCMC sampler allocated to the ratefactor vector ρ during the simulation. The correct number of ratefactors is 3 and has the majority of the density. There is also a substantial proportion for there being 4 or 5 ratefactors. This shows how the model can explore different numbers of ratefactors and still return to provide correct inference. A large number of samples for less than two ratefactors in the vector would result in significantly poorer results. Redundant values will not alter the mosaic structure inferred. Subfigure d) presents three subplots the following: the trajectory along the simulation for the log likelihood, the posterior probability of ν_S and the posterior probability of ν_R. All three parameters show that they are stable and have not altered much during the exploration of the RJMCMC sampling stage.

The branch length vector $\mathbf{w}$ for the correct topology has the posterior mean (to two significant figures) [0.22, 0.16, 0.21, 0.20, 0.20] and for the relative codon rate vector λ (to two significant figures) [0.34, 0.31, 0.34], which are both very close to the data generating processes values. For the Metropolis Hastings simulations,

200 burnin steps followed with 900 sampling steps were given.

4.6.2 Alignment 2

Subsection 4.4.2 describes the sequence alignment used here as synthetically generated data. The alignment has 1500bp with a topology change at site 500, from the tree having taxa 1 and 3 adjacent to having 1 and 2 adjacent. The ratefactor break points are at sites 500, 1000, and 1500, with values 1, 2, and 0.5 in the three sections. Fig 4.5 shows results of simulations in 4 subfigures.

In subfigure a) the posterior probabilities of the topologies are shown along the sites in the alignment. The first subplot represents the topology where the taxa labelled 1 and 2 are adjacent to each other which is the topology used to create the sites in the data sequence alignment from 500 to 1500. The second subplot represents the topology grouping taxa 1 and 3 together and was used as the topology for sites 1 to 500. From the posterior probabilities shown there is a clear break point at site 500 indicating a recombination event. The subfigure b) shows the mean ratefactor values during the simulation along the sites of the sequence alignment with the 50 and 95 percent credibility intervals plotted (eq 4.48). The credibility intervals do not diverge from the mean at the sites around the break points either. The correct values of 1, 2, and 0.5 are clearly seen in the correct regions. Subfigure c) shows the histogram of the number of ratefactors, $\tilde{K}$ that the RJMCMC sampler allocated to the ratefactor vector $\boldsymbol{\rho}$ during the Gibbs simulation. The correct number of ratefactors 3 has the majority of the density and there being a noticeable proportion for the numbers 4 and 5 shows how the model can explore different regions and still return to provide correct inference. Significant number of samples for less than three states would cause the model to infer mistaken results for this alignment. Subfigure d) shows in the three subplots the trajectory along the simulation of the log likelihood, the posterior probability of ν_S and the posterior probability of ν_R. These trajectories indicate convergence.

The branch length vector $\mathbf{w}$ for the topology with taxa 1 and 2 being adjacent is to two significant figures [0.21, 0.16, 0.25, 0.17, 0.19] and for the relative codon rate vector $\boldsymbol{\lambda}$ to two significant figures [0.28, 0.42, 0.29]. For the topology grouping taxa 1 and 3 the branch length vector sampled was $\mathbf{w} = [0.13, 0.27, 0.26, 0.18, 0.17]$ and for the relative rates between the codon positions [0.45, 0.25, 0.3]. Both are

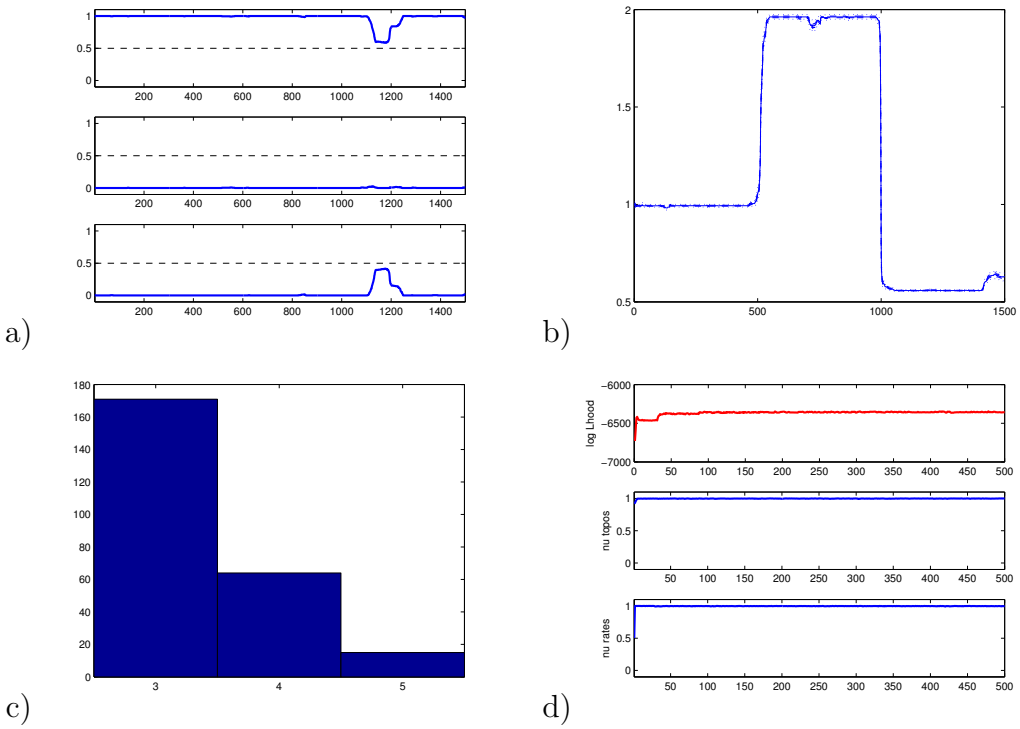


Figure 4.4: Results of the synthetic alignment 1.

Subsection 4.4.1 describes the alignment used as data to produce the results above. Using the phylogenetic tree topology of Fig 4.3 1500bp were generated having three sections of rate heterogeneity. The rate heterogeneity break points are at 500bp and 1000bp. Subfigure a) shows the marginal posterior probabilities of the topologies at each column in the alignment. The topology having taxa 1 and 2 adjacent is represented in the first subplot showing that along the alignment the model allocated close to all the sampled topology states to the correct topology. Subfigure b) shows the sampled posterior mean of the rate factor values allocated along the alignment. The three sections are clearly indicated with values of ρ taking 1, 2, and 0.5 which are what was used to generate the data. The 50 and 95 percent credibility intervals are very close to the mean. Subfigure c) is a histogram of the number of rate factor values that the ratefactor vector, ρ contained during the Gibbs simulation. The majority shows that the correct number of 3 contains the majority of the density. Subfigure d) shows in the three subplots the trajectory of the log likelihood, topology transition parameter ν_S, and rate state transition parameter ν_R which all have stable increasing values.

close to the supplied vectors to the data generating processes. For the Gibbs simulation 200 burnin steps followed by 200 sampling steps were given. For the Metropolis Hastings simulations, step 200 burnin steps followed with 900 sampling steps were given.

4.6.3 Alignment 3

Subsection 4.4.3 describes the sequence alignment parameters used to generate the synthetic data used to test the model's ability of fitting topology break points which do not align with ratefactor break points. The alignment has 1500bp with a topology change at site $\mathbf{y}_t = 750$, from the tree having taxa 1 and 3 adjacent to having taxa 1 and 2 adjacent. The ratefactor break points are at sites 500, 1000, and 1500, with values 1, 2, and 0.5 in the three sections. Fig 4.6 shows the results of simulation in 4 subfigures.

In subfigure a) the posterior probabilities of the topologies are shown along the sites in the alignment. The 3 subplots show for each topology the posterior probability along the sites in the alignment. The first subplot represents the topology where the taxa labelled 1 and 2 are adjacent to each other which is the topology used to create the sites in the data sequence alignment from 750 to 1000. The second subplot represents the topology grouping taxa 1 and 3 together and was used as the topology for sites 1 to 750. From the posterior probabilities shown there is a topology change around the site 750 indicating a recombination event. The subfigure b) shows the mean ratefactor values during the simulation along the sites of the sequence alignment with the 50 and 95 percent credibility intervals plotted (eq 4.48). The two break points separating the three regions of different values of rate heterogeneity are clearly distinct. The credibility intervals do not diverge from the mean at the sites around the break points either. The correct values of 1, 2, and 0.5 are clearly observed. Subfigure c) shows the histogram of the number of ratefactors, $\tilde{K}$ that the RJMCMC sampler allocated to the ratefactor vector $\boldsymbol{\rho}$ during the Gibbs simulation. The correct number of ratefactors 3 has the majority of the density and there being a noticeable proportion for the numbers 4 and 5 shows how the model can explore different regions and still return to provide correct inference. A significant number of samples for less than two states would cause the model to infer mistaken results for this alignment. Subfigure d) shows in the three subplots the trajectory along the simulation of the log likelihood,

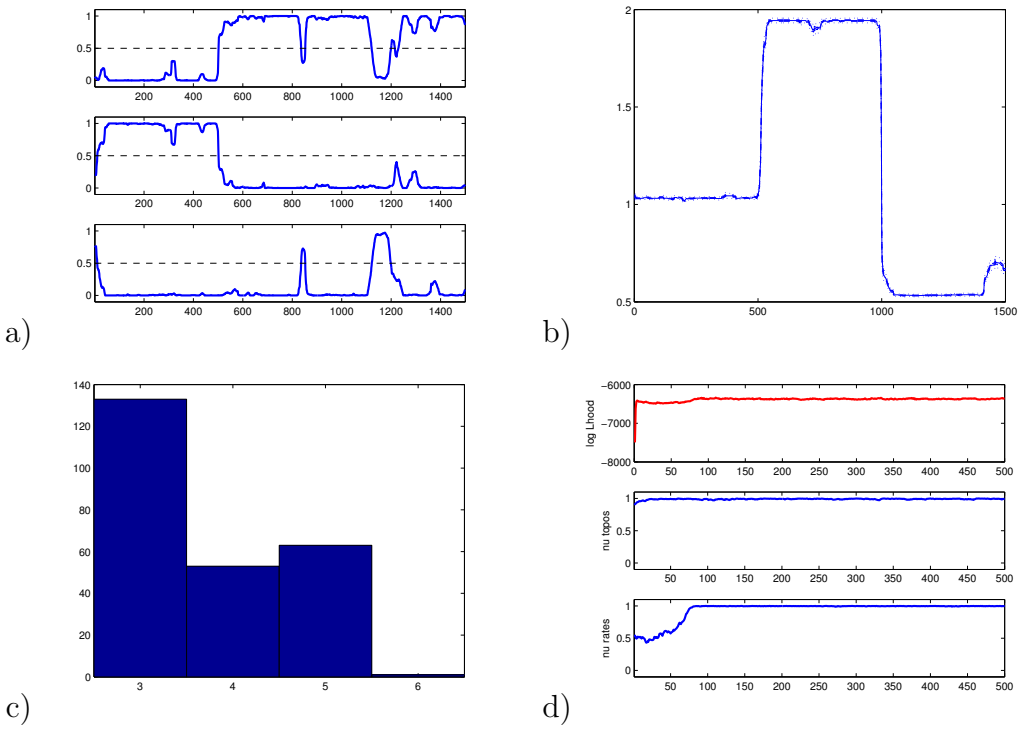


Figure 4.5: Results of the synthetic alignment 2.

Subsection 4.4.2 describes the alignment used as data to produce the results above. Subfigure a) shows the marginal posterior probabilities of the topologies at each column in the alignment. The topology having taxa 1 and 2 adjacent in the phylogenetic tree is represented with the first subplot, the second the tree with taxa 1 and 3 being adjacent. The posterior probabilities along the sites is represented along the alignment. There is a clear topology change at site 500 from the topology of having taxa 1 and 3 adjacent, to the topology having taxa 1 and 2 adjacent. Subfigure b) shows the sampled posterior mean of the rate factor values allocated along the alignment (eq 4.48). The three sections are clearly indicated with values of 1, 2, and 0.5 which are what was used to generate the data. The 50 and 95 percent credibility intervals are very close to the mean. Subfigure c) is a histogram of the number of rate factor values that the ratefactor vector, ρ contained during the Gibbs simulation. The correct number of 3 contains the majority of the density. Subfigure d) shows in the three subplots the trajectory of the log likelihood, topology transition parameter ν_S, and rate state transition parameter ν_R which all indicate convergence.

the posterior probability of ν_S and the posterior probability of ν_R. All three parameters show that they are stable and have not altered much during the exploration of the RJMCMC sampling stage.

The branch length vector $\mathbf{w}$ posterior mean for the topology with taxa 1 and 2 being adjacent is to two significant figures [0.20, 0.23, 0.15, 0.20, 0.21] and for the relative codon rate vector λ to two significant figures [0.34, 0.38, 0.28]. For the topology grouping taxa 1 and 3 the branch length vector sampled was $\mathbf{w} = [0.14, 0.25, 0.22, 0.20, 0.19]$ and for the relative rates between the codon positions [0.36, 0.30, 0.34]. Both are close to the supplied vectors to the data generating processes. For the Gibbs simulation 200 burnin steps followed by 200 sampling steps were given. For the Metropolis Hastings simulations, step 200 burnin steps followed with 900 sampling steps were given.

4.6.4 Short alignments

In this section simulations with alignments of fewer base pairs are examined. All the simulations of this section had for the Gibbs simulation 250 burnin steps and 250 sample phase steps. Other simulation lengths were also tested but minor differences to convergence were made by giving an increase in iterations. Degradation of the quality of the results occurs most commonly when the burnin phase is less than 150 iteration for the burnin and the sample phase.

Figure 4.7 shows the results from the model run on an alignment of 300bp. The alignment was produced with 2 recombination events and three different regions of rate heterogeneity whose changepoints occur at sites 100 and 200. Subfigure d) is a diagram showing the structure of these features along the alignment. We expect to see topology changes at these points from topology 1 to 2 and then to 3. Subfigure a) shows the posterior probabilities along the sites for the three topologies. The signal is not very stable along any of the topologies which is due to the lack of sufficient data for there to be less uncertainty. However, there is a clear indication that the model is being affected by the change in the topology along the alignment. Subfigure b) shows the ratefactors with the credibility intervals of 50 and 95% (according to eq 4.48). The 3 rate states are clearly distinguishable and have the changepoints close to the sites of 100 and 200. Around those site the credibility intervals are wide as the allocation of the rate states can shift sites placing different rate factor values there different from previous

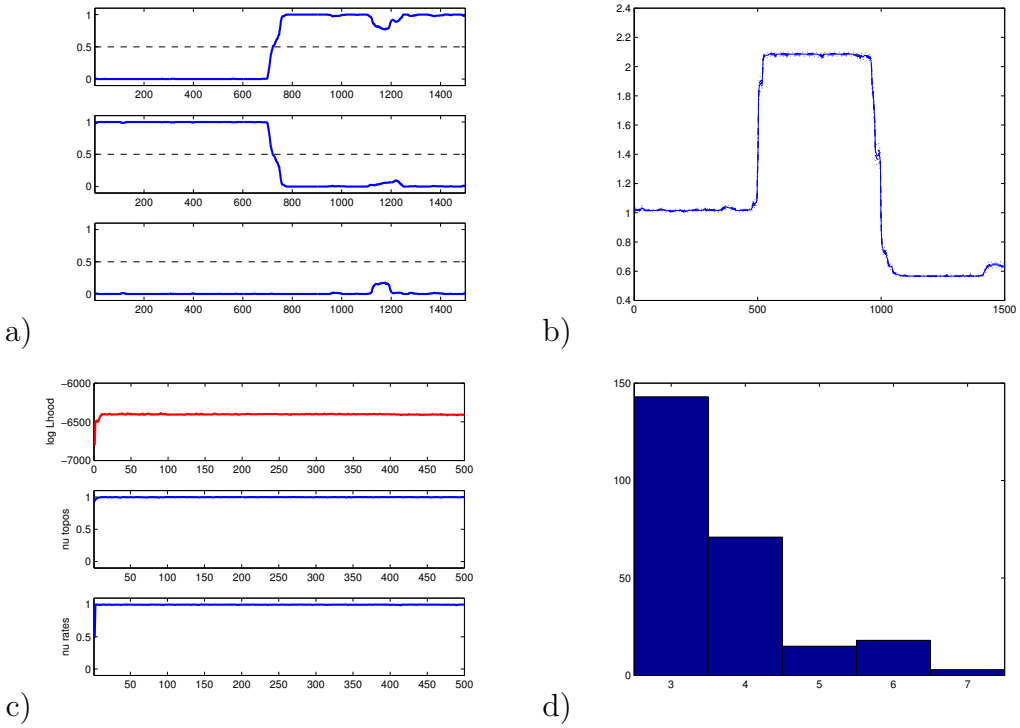


Figure 4.6: Results of the synthetic alignment 3.

Subsection 4.4.3 describes the alignment used as data to produce the results above. The phylogenetic tree topology grouping taxa 1 and 3 is used for producing the first 750bp of the alignment, and the rest of the 1500bp from the topology grouping taxa 1 and 2. There are three generated sections of rate heterogeneity. The rate heterogeneity break points are at 500bp and 1000bp, with rate factors of 1, 2 and 0.5. Subfigure a) shows the posterior probabilities of the topologies at each column in the alignment for each topology in a subplot. The topology having taxa 1 and 2 adjacent in the phylogenetic tree is represented with the first subplot, the second the tree with taxa 1 and 3 being adjacent. There is a topology change across the corresponding subplots around the site 750 is correct. Subfigure b) shows the sampled posterior mean of the rate factor values allocated along the alignment (eq 4.48). The three sections clearly indicated with values of 1, 2, and 0.5 which are the values used to generate the data. The 50 and 95 percent credibility intervals are very close to the mean. Subfigure c) is a histogram of the number of rate factor values that the ratefactor vector, ρ contained during the Gibbs simulation. It shows that the correct number of 3 contains the majority of the density. Subfigure d) shows in the three subplots the trajectory of the log likelihood, topology transition parameter ν_S, and rate state transition parameter ν_R along the iterations of the simulations have stable increasing values.

samples. Subfigure c) shows the histogram of the number of ratefactors the RJMCMC sampler produced during the simulation. For the 3 topologies the branch length vectors uncovered were: [0.17;0.03;0.40;0.29;0.11],[0.24;0.05;0.23;0.30;0.17] and [0.41;0.07;0.11;0.2;0.2]. The relative codon rate vectors for the topologies sampled were: [0.11;0.81;0.08], [0.43;0.39;0.18] and [0.24;0.10;0.66].

Figure 4.8 shows the results from the model run on an alignment of 300bp. The alignment was produced with a recombination event at site 150 and three different regions of rate heterogeneity whose changepoints occur at sites 100 and 200. The difference from figure 4.7 is that the changepoints for topology changes do not lie on the same sites as the change points for the ratefactors. Subfigure d) is a diagram showing the structure of these features along the alignment. We expect to see topology changes at these points from topology 1 to 2 in the center of the alignment. Subfigure a) shows the posterior probabilities along the sites for the three topologies. The signal is stable enough to clearly determine the correct topologies along the alignment and the change point at site 150 is clearly seen. Subfigure b) shows the ratefactors with the credibility intervals of 50 and 95% (according to eq 4.48). The 3 rate states are not all clearly distinguishable as the second changepoint expected at site 200 is not strong enough to indicate a new state. This can be due to the fact that the topology changes can assist the model to find stronger incentive for the creation of a new ratefactor state. Subfigure c) shows the histogram of the number of ratefactors the ratefactor vector contained during the simulation. It is evident that the model does not hold often more than 2 ratefactors. Longer simulations were run to examine whether this lack of a changepoint at site 200 is due to convergence or lack of data, and differences were not noted with a Gibbs burnin and sampling phase of 400 and 700 respectively. For the 3 topologies the branch length vectors uncovered were: [0.22;0.09;0.38;0.05;0.26], [0.18;0.10;0.30;0.10;0.31] and [0.17;0.18;0.07;0.15;0.43]. The relative codon rate vectors for the topologies sampled were: [0.27;0.40;0.33], [0.30;0.40;0.30], and [0.22;0.43;0.35].

Figure 4.9 shows the results from the model run on an alignment of 750bp. The alignment was produced with recombination events at sites 250, and 500. There are 3 different regions of rate heterogeneity whose changepoints occur at sites 250 and 500. The purpose is to examine the effect of having an alignment of the same format as figure 4.7 when it is longer. Subfigure d) is a diagram showing the structure of these features along the alignment. We expect to see topology

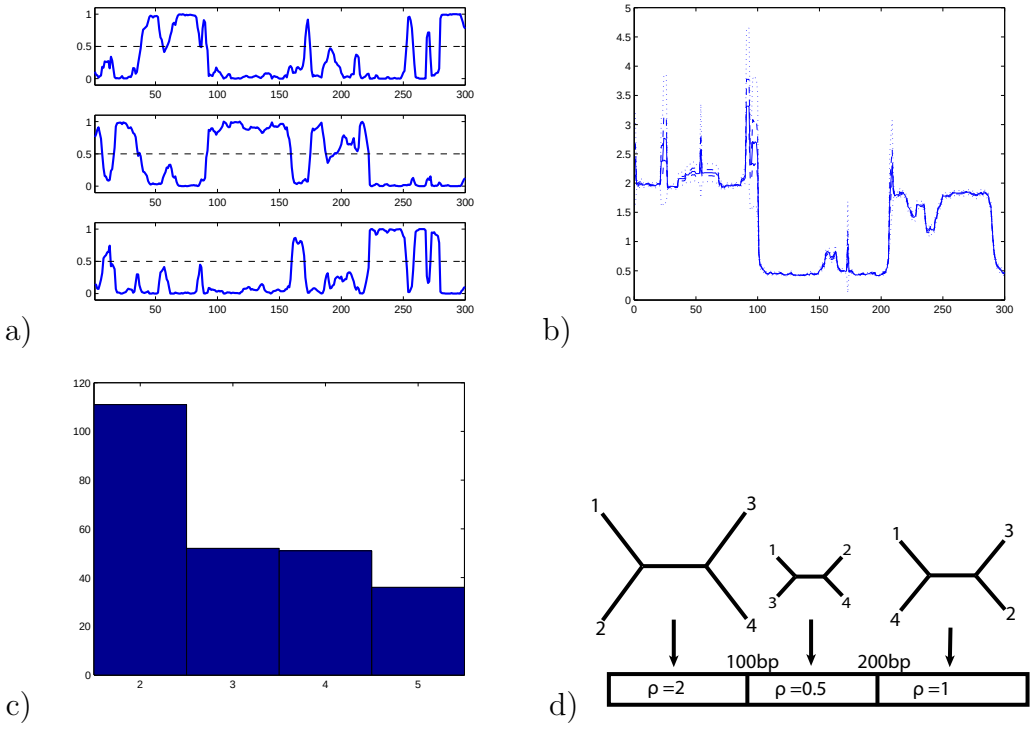


Figure 4.7: 300bp alignment with 3 regions of rate heterogeneity and 2 recombination events.

A synthetically produced sequence alignment 300bp long is generated. There are 3 recombination events and three regions of rate heterogeneity with change points at sites 100 and 200. Subfigure d) shows a diagram of the structure of the alignment. Subfigure a) shows the marginal posterior probabilities of the topologies along the sites of the alignment where the topology structure of the data generating process can be seen with the addition of noise. Subfigure b) shows the mean of the posterior samples of the ratefactors with the 50 and 95 percentile credibility intervals. There is large uncertainty around the sites of the changepoints but the structure of the rates is uncovered. Subfigure c) shows the histogram of the number of ratefactors allocated from the RJMCMC scheme.

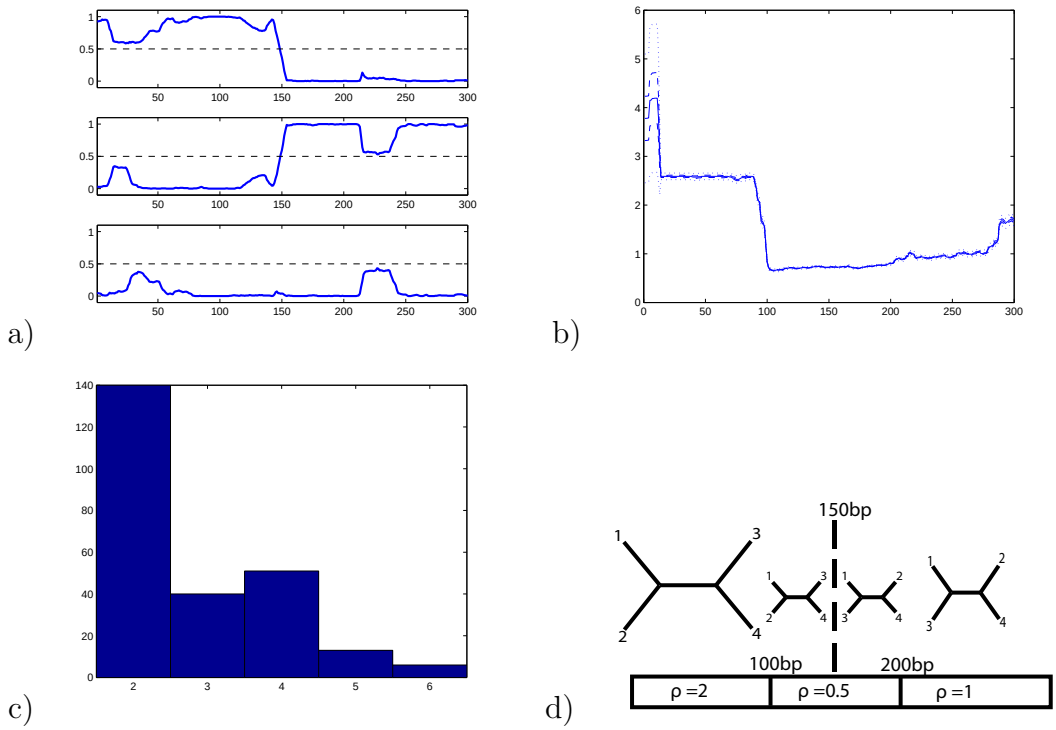


Figure 4.8: 300bp alignment with 1 recombination event and 3 regions of rate heterogeneity.

A synthetically produced sequence alignment 300bp long is generated. There is a recombination event at site 150 and three regions of rate heterogeneity with change points at sites 100 and 200. Subfigure d) shows a diagram of the structure of the alignment. Subfigure a) shows the marginal posterior probabilities of the topologies along the sites of the alignment where the topology structure of the alignment shows the break point in the middle of the alignment. Subfigure b) shows the mean of the posterior samples of the ratefactors with the 50 and 95 percentile credibility intervals (dashed and dotted lines respectively). It can be seen that the changepoint at site 200 is not present. Other simulations show a noticeable gradient towards the end of the alignment but none show clear distinguishable change point. Subfigure c) shows the histogram of the number of ratefactors allocated from the RJMCMC scheme. The model's prior outweighs the likelihood of greater number of ratefactors to apply another changepoint to the alignment.

changes at these points from topology 1 to 2 at site 250 and from 2 to 3 at site 500. Subfigure a) shows the posterior probabilities along the sites for the three topologies. The signal is stable enough to clearly determine the correct topologies within the expected regions. Subfigure b) shows the ratefactors with the credibility intervals of 50 and 95% (according to eq 4.48). The 3 rate states are all clearly distinguishable. Subfigure c) shows the histogram of the number of ratefactors the RJMCMC sampler produced for the ratefactor vector during the simulation. The quality of the results reveals the limit to how much data is needed for correct inference in such a situation. For the 3 topologies the branch length vector posterior means uncovered were: [0.19;0.18;0.16;0.28;0.19], [0.23;0.22;0.1;0.29;0.17] and [0.23;0.26;0.12;0.23;0.16]. The relative codon rate vector posterior means for the topologies sampled were: [0.38;0.31;0.31], [0.38;0.29;0.33], and [0.22;0.51;0.27].

Figure 4.10 shows the results from the model run on an alignment of 750bp. The alignment was produced with a recombination events at site 375. There are 3 different regions of rate heterogeneity whose changepoints occur at sites 250 and 500. The purpose is to examine the effect of having an alignment of the same format as figure 4.8 when it is longer, and see whether the model still has the inability in not finding the third region of rate heterogeneity. Subfigure d) is a diagram showing the structure of these features along the alignment. We expect to see a topology change at site 375, from topology 1 to 2, and changepoints in the ratefactor allocation states at sites 250 and 500. Subfigure a) shows the posterior probabilities along the sites for the three topologies. The signal is stable enough to clearly determine the correct topologies within the expected regions. There is a presence of noise but not strong enough to infer an incorrect topology and multiple runs of the simulation did not remove this problem. Subfigure b) shows the ratefactors with the credibility intervals of 50 and 95% (according to eq 4.48). The 3 rate states are all clearly distinguishable. There is a spike in the value at the 500bp changepoint site. The credibility intervals are wide as well. This is an improvement over the analogous alignment of 300bp which did not produce a clear changepoint in the rate state. For limited data not having a crisp breakpoint is anticipated. Subfigure c) shows the histogram of the number of ratefactors the RJMCMC sampler produced during the simulation.

For the 3 topologies the branch length vectors uncovered were: [0.22;0.14;0.25;0.20;0.18], [0.24;0.12;0.23;0.21;0.20] and [0.19;0.11;0.10;0.36;0.25]. The relative codon rate vectors for the topologies sampled were: [0.34;0.31;0.34],

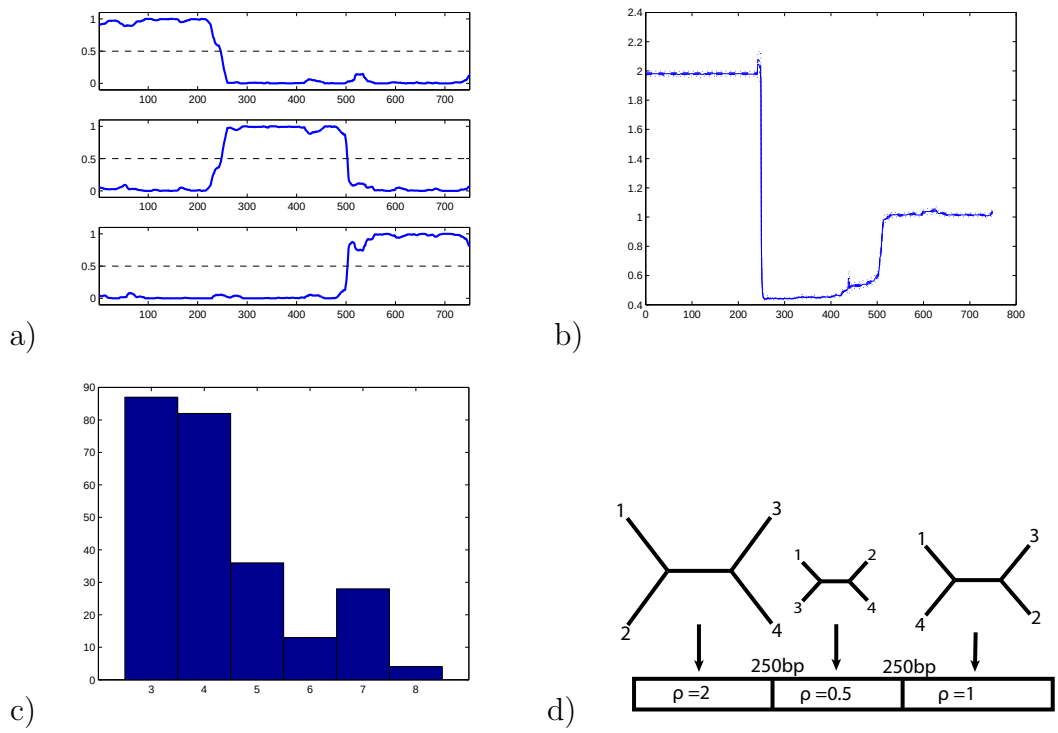


Figure 4.9: 750bp alignment with 2 recombination events and 3 regions of different rate heterogeneity.

A synthetically produced sequence alignment 750bp long is generated. There are recombination events at sites 250 and 500, and the three regions of rate heterogeneity have change points occurring at the same sites. Subfigure d) shows a diagram of the structure of the alignment. Subfigure a) shows the marginal posterior probabilities of the topologies along the sites of the alignment where the topology structure of the alignment shows the break points at sites 250 and 500 clearly. The topologies in the respective regions have stable majorities of the posterior distributions along the sites. Subfigure b) shows the mean of the posterior samples of the ratefactors with the 50 and 95 percentile credibility intervals. The changepoints are clear and at the correct sites. The ratefactor values are also correct with tight credibility intervals. Subfigure c) shows the histogram of the number of ratefactors allocated from the RJMCMC scheme.

[0.33;0.34;0.33], and [0.27;0.30;0.44].

4.6.5 Note about the results of the ratefactors along the sites

From the figures of the results of the simulations presented in this section there is a feature commonly observed which is not explained by the data generating process. This feature is in the subfigures b) of the figures and requires an explanation. In these subfigures the mean of the ratestate values sampled along the sites of the alignment is plotted as calculated in eq 4.48. In the figures 4.4 to 4.8 it can be seen that there is a change in the mean plotted value shortly before the last sites in the alignment. In the case of figure 4.7 this aligns with a topology change, but in the rest of the simulations a small increase can be observed.

There is a geometric prior on the length of a segment given the way that the state transitions are modelled. The probability for a segment length N is $(1-\nu)^{N-1}\nu$. Shorter segment lengths have greater support in the absence of enough data. Ratestate changes into an inaccurate ratefactor value will not create a large penalisation when applied to a small number of sites (ie. towards the end of the alignment). As the sampling procedure proposes moves into alternative rate state values, incorrect ones can more likely be accepted in this region. The increase rather than decrease is because in these simulations the alternative ratefactor values available are all of a greater value. An average of the occasionally greater ratefactor values and the correct value creates a slight increase.

4.7 Discussion

The generalised FHMM of Mantzaris and Husmeier (2009) was extended to include the transdimensional sampling for the rate factors along the alignment. The work of this chapter addressed the limitation in that the number of rate factors had to be defined beforehand (the size of the rate factor vector ρ was fixed during the simulations). The improved phylogenetic FHMM developed in this chapter includes the sampling of the branch lengths $\mathbf{w}$, the relative vector for the codon level of rate heterogeneity λ, and now the RJMCMC scheme introduced to sample the number of number of ratefactors $\tilde{K}$ for the ratefactor vector ρ. As shown with the synthetic sequence alignments the model is able to find the regions of rate heterogeneity along the sequence alignment applying appropriate

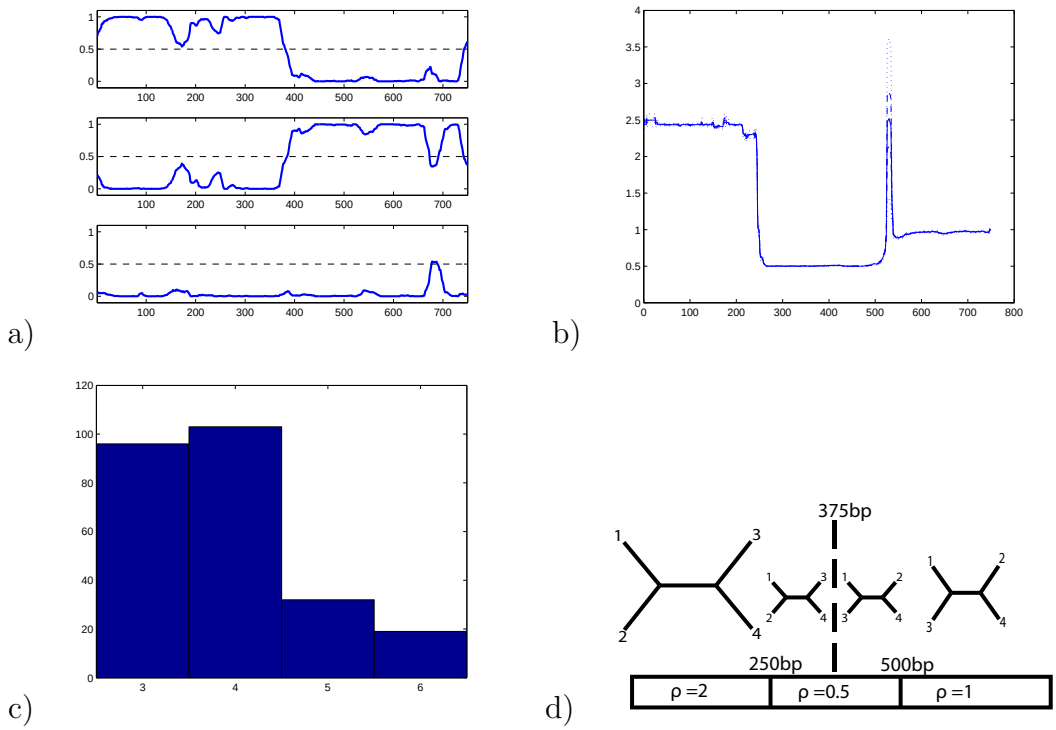


Figure 4.10: 750bp alignment with 1 recombination event and 3 regions of different rate heterogeneity.

A synthetically produced sequence alignment 750bp long is generated. There is a recombination event at site 375, and three regions of rate heterogeneity change points occur at the same sites. Subfigure d) shows a diagram of the structure of the alignment. Subfigure a) shows the posterior probabilities of the topologies along the sites and the expected change point for the correct topologies is seen in the center about site 375. There is some noise in the signal and running the program for longer or reproduced synthetic data does not remove sporadic disturbances in the signal. Subfigure b) shows the mean of the posterior samples of the ratefactors with the 50 and 95 percentile credibility intervals. The changepoints are crisp and at the right sites. The ratefactor values are correct with tight credibility intervals. The changepoint at site 500 has a large sporadic spike and a wide credibility interval. This does not always appear but is a feature that is not uncommon at the changepoints when there is a lack of data. This is one more example of how the structure of this alignment where the topology changes are not inline with the different regions of rate heterogeneity creates difficulty for the model. Subfigure c) shows the histogram of the number of ratefactors allocated from the RJMCMC scheme.

break points and sampling the correct number of ratefactors. The credibility intervals for the sampled parameter values are close to the sampled means in the simulations. Convergence was obtained for the sampled parameters according to the Gelman and Rubin potential scale reduction factors with values below 1.2. The synthetic DNA sequence alignments had their underlying parameters used in the data generating process found with the break points along the sites as well. Topology changes were able to be inferred independently of the ratefactors.

The model with the extension copes with the increased complexity in the presence of a sufficient amount data. Without sufficient data the uncertainty in the estimates produces results with many oscillations between topology estimates and an inaccurate number of inferred ratefactors. With sufficient data the alignments with many features did not display excessive uncertainty. Therefore it can be concluded that the method copes with the improvement made in the presence of sufficient data supplied for the break points of either the topologies or ratefactors.

Chapter 5

Application to Neisseria

This chapter deals with the analysis of a real world DNA sequence alignment of *Neisseria* and the presence of inter species recombination. This data set (the *Neisseria* alignment) was extracted and initially analysed in Zhou and Spratt (1992). The genes extracted from the genome are the argF, fbp, and recA genes. The sequences can be found in the EMBL database with pubmed id number 140654. The work of Husmeier (2005), which introduces the phylogenetic FHMM, uses a subset of this *Neisseria* data set by choosing 4 sequences in creating an alignment. (The phylogenetic FHMM is discussed in this thesis in subsection 1.9.4.)

The 4 sequences chosen in Husmeier (2005) were based on the work done in Zhou and Spratt (1992) whose findings revealed that there is recombination in the argF gene. In both these works the authors use a simpler inference process to determine the recombination break points compared to the method discussed in this chapter. Because of the limitations of the earlier models, it is possible that there are other recombination break points between other strains which can be found via the improved PFHMM developed in this thesis. The improved phylogenetic FHMM (PFHMM or phylo-FHMM) is too computationally demanding to analyse directly all of the sequences submitted by Zhou and Spratt (1992). The number of sequences must be reduced to restrict the size of the topology search space. Other methods are employed to assist in the process of choosing alignments of 4 sequences to be analysed. These methods help 'prune' unnecessary sequences from the alignment. A brief overview of the process of pruning the alignment is given here, and the different methodologies applied are mentioned here as a primer.

The first method applied to the sequence alignment is that of phylogenetic net-

works described in section 5.1.2. It constructs a phylogenetic network from the complete submitted data set of *Neisseria* sequences. The phylogenetic network has similarities to a phylogenetic tree with the difference that it displays indications of evolutionary events which disagree with a single tree topology (events such as recombination), and that cycles are introduced. The method delivers a very coarse indication of which strains in the given data set contain possible recombination break points and which strains are closely related. Later in this chapter it is demonstrated how using the network constructed from the sequences a simplification is made to reduce the number of strains of meningitidis used.

In the following subsection 5.1.2 phylogenetic networks are then applied to the reduced set of strains. All possible phylogenetic networks of 4 DNA sequences are generated and presented. The set of alignments which indicated the largest presence of possible recombination between them are identified.

The DSS statistic is then introduced in section 5.2 and is applied in subsection 5.2.2. This method produces more detailed information about possible recombination break points along alignments. Guided by the results of applying DSS to the reduced set of alignments, the number of alignments of 4 sequences is further reduced. With the same motivation BARCE is then discussed in section 5.3 and applied in subsection 5.3.2 to the alignments highlighted from DSS. The improved analysis offered by BARCE will then be compared to the simulation results produced by the improved PFHMM subsequently.

For completeness, on the *Neisseria* data set the DSS and BARCE highlighted alignments are also used with the different alignments chosen in Husmeier (2005) and Zhou and Spratt (1992). These results are presented in subsection 5.4.1. This gives a comparison of the recombination break points that arise from each of the methods of analysis. Calculating PFHMM results is computationally expensive, so to demonstrate the use of the improved PFHMM a selection of alignments are chosen from the BARCE simulations and are then processed with the improved PFHMM, shown in section 5.4.2.

Finally, a conclusion section discusses the work done in this chapter. The methodological and biological conclusions that can be derived from the simulations are given. An overall conclusion about the investigation is made and an avenue for future work is proposed.

5.1 Phylogenetic Networks

From the available phylogenetic network methods, in this work *SplitsTree4*, from Huson and Bryant (2006), is chosen to produce phylogenetic networks. This method is chosen because it is well established, easily tractable and well supported in software. This choice is not significant to the findings of this work. The paper, Huson and Bryant (2006) provides a detailed introduction to the method and the underlying theory. The authors discuss the use of these networks as a preliminary step to tree-based analysis which is how the method is used in this thesis for selecting sequence alignments. To summarise the process, alignments are produced from the publicly available Neisseria sequences with ClustalW, and then phylogenetic networks are produced from the alignments using SplitsTree4.

5.1.1 Definition

Phylogenetic networks attempt to model evolutionary events which cannot be described by a single tree. A tree topology is appropriate when the sequences have evolved via point mutations alone, but inappropriate for expressing features such as a recombination event. These networks display the set of possible tree topologies which there is evidence for in the data. This is done by examining each column of the alignment independently and producing a phylogenetic tree for the column, with a non-probabilistic method. These trees are then combined into a single network in an additive manner. In combining trees into a network, edges which overlap between 2 trees are mutually reinforcing and so have their lengths increased in a cumulative manner. In cases where edges do not overlap between trees, both possibilities must be represented in the phylogenetic network. This is achieved by introducing 2 auxiliary nodes with 2 pairs of parallel branches equal in size to the conflicting branches (conflicting in that they do not support a single bifurcating tree). This introduction creates a trapezoid from where there was originally a single line. Phylogenetic networks may be simplified by collapsing conflicted nodes, which if continued will ultimately collapse to a tree structure. When collapsing conflicted nodes, removing a pair of edges and their auxiliary nodes simply restores one of the original network topologies used to produce the split. In these networks, as in phylogenetic trees, the lengths of all the edges are proportional to the expected number of substitutions between the taxa. There are frequently many splits in a phylogenetic network. Visualisation of a phylo-

genetic network lends itself easily to interpretation, any splits are represented as trapezoidal areas and the magnitude of the conflicts between network topologies is seen from the sizes of the trapezoidal regions.

A phylogenetic network of a sequence alignment gives indications of recombination events (horizontal transfer of genetic material), gene duplication or loss, and hybridisation that may have occurred in the genetic history of the taxa. All these events are non-linear, but in this work the focus is on recombination because it is the natural point for continuation of earlier published work. There is much scope for improvements of the SplitsTree4's method of discovering recombination as phylogenetic networks do not find the sites for the break points of these non-linear evolutionary events. In general not only is the location of these events not represented by phylogenetic networks, these networks also do not explicitly provide indications for which of the possible non-linear evolutionary events might have occurred.

Figure 5.1 shows an image of a phylogenetic network produced from the set of DNA sequences containing the argF, fbp, and recA genes of *Neisseria*. This was published by Zhou and Spratt, and the sequences can be found in the EMBL database with pubmed id number 140654. The publication, Zhou and Spratt (1992), identifies a possible recombination event in the argF gene. This can be seen in the splits between the sequences with large trapezoids indicating large conflicting signals between sequences in regions on the network. The naming on the phylogenetic network (from the scheme in EMBL) is simplified by using an abbreviation. The corresponding accession numbers of the sequences are replaced by the first letter of the strain's name and the last 2 digits of the identification number. The groups are 'G' for Gonorrhoeae, 'M' for Meningitidis, 'C' for Cinerea, 'P' for Polysaccharea, 'L' for Lactamica, 'F' for Flavescens and 'Mu' for Mucosa. From the network it can be seen that there is a substantial amount of area in the trapezoidal netting involving the strains 'G', 'P', and 'L'. Recombination detection will then likely involve these strains. The 'M' strains have very small distances between them, which is reasonable given that they are from the same family of strains. They can then be considered identical for the purpose of detecting recombination and have a representative strain used in place of the set of strains. This simplification will greatly reduce the set of possible alignments of 4 sequences needed for using BARCE and the improved PFHMM.

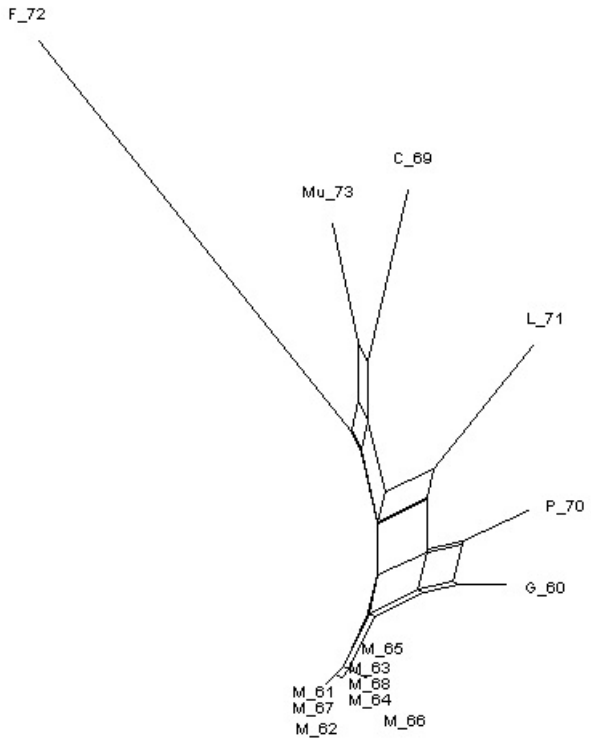


Figure 5.1: Figure of phylogenetic network produced from sequence of *Neisseria* with possible recombination in the argF gene

The phylogenetic network produced from the sequence alignment of *Neisseria* containing the argF, fbp and recA genes. The phylogenetic network has abbreviated labels. . In the diagram G stands for *Neisseria Gonorrhoeae*, M for N.*meningitidis*, C for N.*cinerea*, P for N.*polysaccharea*, L for N.*lactamica*, F for N.*flavescens* and Mu for N.*Mucosa*. Strains G, P, and L produce large conflicting signals for a single topology (indicated by large trapezoidal regions). This highlights them as candidates for detecting recombination. The group of M strains can be reduced to a single representative strain. There is little indication for recombination (conflicting signals for a single topology) between the family of M strains.

5.1.2 Application of Phylogenetic Networks

The next step in the analysis is to use phylogenetic networks to analyse the Neisseria alignment by producing alignments of 4 sequences. The results will provide evidence as to which alignments have the largest indication of recombination. These alignments will then be used with a more complex method for identifying recombination. There are 7 families of strains (as given by the results of figure 5.1) to be grouped into alignments of 4 sequences, and a full search of the possibilities requires $\binom{7}{4} = 35$ networks to investigate.

Figure 5.2 shows the results of this investigation with phylogenetic networks of four sequences. Conflicting signals in the networks can be seen as trapezoidal regions. The size of the trapezoidal regions indicates the degree to which there is evidence of a lack of support for a single phylogenetic tree. The images of the networks are not rescaled so their sizes can be interpreted by the reader for obtaining an indication of the expected number of mutations along the branch lengths. It can be seen that the networks with the largest enclosed trapezoidal areas are in subfigures 15, 16, 17, 18, 19 and 20. These subfigures correspond to: (M,Mu,L,G), (M,Mu,P,G), (M,C,L,P), (M,C,L,G), (M,C,P,G), and (M,L,P,G). The commonality of the strain M in all these highlighted alignments, in combination with the absence of such areas in alignments which do not include the strain M provides strong evidence that a recombination event can be anticipated when strain M is included in an alignment. Using this evidence to propose strain M as a reference strain, further analysis shows that the strains G,P and L are sufficient in combination with strain M to show evidence of recombination under analysis of the 4 sequence phylogenetic networks. In summary recombinant strains with respect to M can be detected in any network that also includes strains G, P or L. Having determined the probable recombinant strains, this reduced set of alignments will be used with more complex methods to obtain a more accurate picture of the possible recombination events in these sequences.

5.2 DSS: Difference of Sums of Squares method

The difference of sums of squares method (DSS statistic eq 5.2) is used to infer possible recombination events in sequence alignments. The method uses a rolling window to segment the sequence alignments and then examines the consistency of

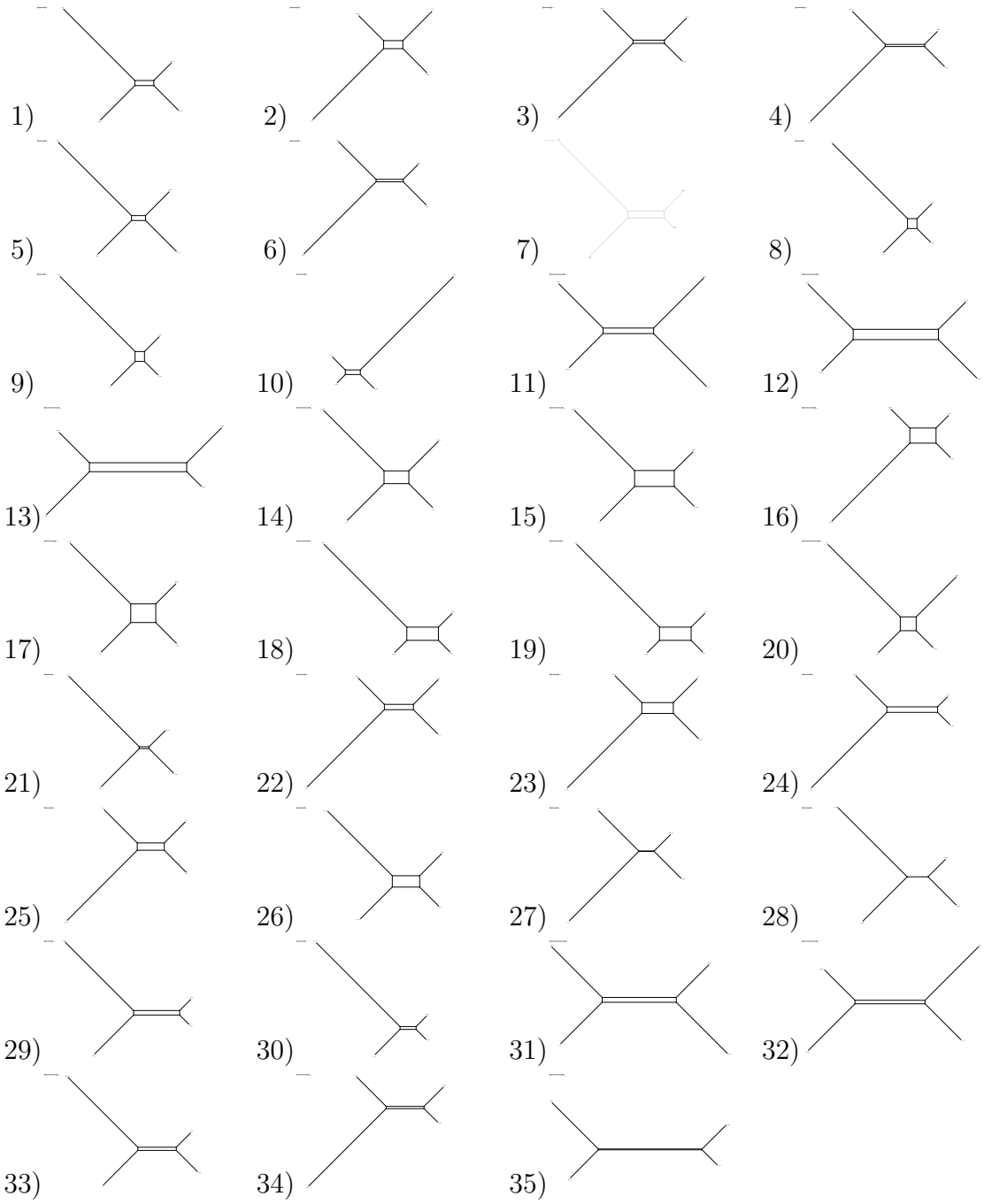


Figure 5.2: The phylogenetic networks of 35 possible alignments of the argF gene of *Neisseria*

Phylogenetic networks are produced from alignments of 4 sequences. From the total set shown here 15, 16, 17, 18, 19 and 20 show the largest rectangular regions (proportional to the conflicting signal for a single tree topology). These alignments will be used subsequently with a more complex methodology for inferring recombination. These alignments highlighted correspond to: (M,Mu,L,G), (M,Mu,P,G), (M,C,L,P), (M,C,L,G), (M,C,P,G), and (M,L,P,G). The abbreviation used here is described in the caption of figure 5.1.

the scores of candidate phylogenetic trees within the segments. Large differences within a region indicate increased likelihood of topology break points possibly caused by recombination events having occurred within that region. In calculating the DSS statistic the windowed section of the sequence alignments is divided into 2 halves and the first half is used to estimate a reference phylogenetic tree. Using this reference tree a goodness-of-fit score is then computed for both the first and the second half of the window. The difference between these scores gives the DSS statistic which is used to determine whether a recombination event is present or not. The window sizes can be configured and affect the sensitivity of the DSS statistic. Much of the background to the method is described in McGuire and Wright (2000), and is used in the TOPALi software presented in that same publication. Generally a window size of approximately 200-500 base pairs is used, and a window size of 400 is recommended by the original authors. In this work, following preliminary validating experiments we find that the suggested window size of 400 base pairs is adequate as recommended. Figure 5.3 depicts this method. Because of the computational cost maximum likelihood is not used for inferring the reference tree on the first (left) half of the window, instead a distance metric is used. Subsection 1.4.1 describes distance methods and the shortcomings they have.

5.2.1 Definition

For the equation below, d_i is used to denote the set of pairwise distances of the sequence alignments in the left side of the window. The reference tree is generated from these pairwise distances using the neighbour joining algorithm. $\tilde{d_i}$, is the set of pairwise distances for the right side of the window. The reference tree has an associated set of pairwise distances e_i, with which the distance to the sets d_i and $\tilde{d_i}$ are compared. From the two halves of the window, 2 goodness-of-fit scores are obtained; SS_l and SS_r for the left and right halves respectively. The equation is:

$$SS_l = \Sigma_i (d_i - e_i)^2, SS_r = \Sigma (\tilde{d_i} - e_i)^2 \tag{5.1}$$

and the DSS statistic is the absolute difference between these scores:

$$DSS = |SS_r - SS_l|. \tag{5.2}$$

Every windowed region consists of two halves and a generated reference tree. If the reference tree has a similar goodness-of-fit score for both halves then the

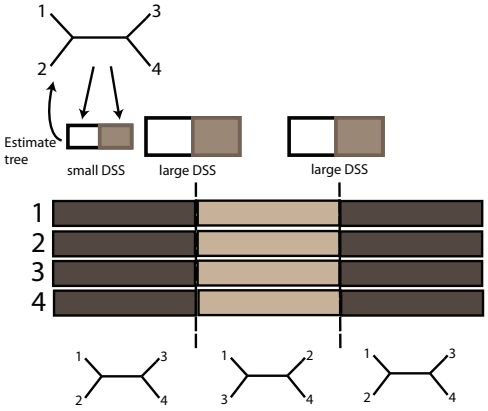


Figure 5.3: The DSS method for detecting regions of recombination

A pictorial description of the DSS method is shown in the figure. The sliding window is moved across the sequence alignment. At each step a reference tree is computed from the left side and distances to the reference tree are compared for the left and right halves. The DSS statistic (eq 5.2) is computed to measure the difference between these 2 regions. In regions where there is no recombination the statistic will give a low value and when the window is centered around the point where there is recombination the statistic will give a high value. Adapted from figures of Husmeier *et al.* (2005a).

reference tree was a good fit over the window and the DSS statistic will be low. Conversely if the reference tree fails to adequately fit both halves then the DSS statistic will be high. Thus the method delivers an approximate location for the presence of recombination events between the first and second half of the window as inferred by the DSS statistic. When the DSS statistic is large (the absolute value of the difference is taken) it is an indication that a recombination event has taken place near the window's centre. To be acceptable the values of DSS statistics need to be tested for significance, to do this parametric bootstrapping is used to compute the distribution of the DSS peaks under the null hypothesis where there is no recombination. The DSS statistics are compared to the bootstrapped values to identify significant peaks. The bootstrap method proceeds by calculating the DSS statistics for perturbed sequence alignments generated by sampling with replacement the columns of the original sequence alignment data. For all sites along the sequence alignment the significance of the DSS statistic is calculated for the original (unperturbed) sequence alignment against the sampled distribution of the DSS statistics for the perturbed sequence alignments via bootstrapping. In this work a significance of 95 percent is required to reject the null hypothesis that no recombination event occurred.

The main drawback of this method is that by using a distance based method structural information is being lost, as discussed in subsection 1.4.1. The uncertainty in the estimation of a reference tree is also not captured by the method. Additionally the choice of the size of the window plays a large role in the sensitivity of the method, smaller window sizes will produce more DSS peaks than larger windows but also a larger values of the bootstrapping confidence interval. For the intended purpose of obtaining a more refined subset of alignments from the set found using phylogenetic networks (5.1.2), DSS is a fast and convenient method.

5.2.2 Application of the DSS statistic

Topali version 1 is used for computing the DSS statistic along the alignments. For DSS, a window of 400 base pairs is chosen as advised by the original authors who use this length in the paper presenting the method. The sequence alignment used in this work is relatively short, and there may not be enough data to support the use of a complex nucleotide substitution model. The Jukes-Cantor model of

nucleotide substitution is used by the authors and this model is our chosen model of nucleotide substitution as well.

The previous analysis shown in figure 5.2 was performed with phylogenetic networks, and the highlighted alignments (detailed in the caption) are used as the initial sequence alignment set for this analysis. In addition to the initial set of 6 alignments, two extra alignments are included; (M,Mu,C,P) and (F,Mu,L,G). These are included to improve the repertoire; the first additional sequence alignment does not include the strain G which is otherwise represented in the selection almost without variation, and the second one is added to include strain F which is not present at all in the rest of the alignments.

From the previous analysis shown in figure 5.2 a subset of the alignments whose results are shown are chosen to be analyzed with DSS. The full set of alignments chosen to be analysed with DSS is (referring to the subfigures of figure 5.2): (M,Mu,C,P) 12), (M,Mu,L,G) 15), (M,Mu,P,G) 16), (M,C,L,P) 17), (M,C,L,G) 18), (M,C,P,G) 19), (M,L,P,G) 20), and (F,Mu,L,G) 23). The numbers and abbreviations are shown in the caption of figure 5.1.

Figure 5.4 shows in subfigures a) to h) the results of these DSS simulations (using TOPALi) on the alignments (M,Mu,C,P) 12), (M,Mu,L,G) 15), (M,Mu,P,G) 16), (M,C,L,P) 17), (M,C,L,G) 18), (M,C,P,G) 19), (M,L,P,G) 20), and (F,Mu,L,G) 23). The subfigures b) and e) look identical at the first glance but do have small differences. The red dotted line is from the bootstrap 95 percentile which is a reference point for the null model of no recombination. Values above the bootstrap value are indications that the DSS statistic (eq 5.2) for the data shows a sufficiently strong signal for a non-homogeneous region. From the diagrams produced by TOPALi, the subfigures a, b, d and e show significantly larger indications for heterogeneous regions than the other 4. This subset is therefore the focus of further analysis. The other 4 subfigures do show occasional peaks above the threshold, but the irregularity of the occurrences do not provide strong support for recombination. The reason for discarding these sporadic peaks is that the recombination event would have to be of a significant length to be detectable using these methods, and the peaks seen on subfigures c,f,g and h are too brief to be distinguished from sampling noise. Since the 95 percent confidence interval for the bootstrap is taken, 1 in 20 sites can be expected to pass the threshold, and alignments showing at least 50 sites above the threshold are taken.

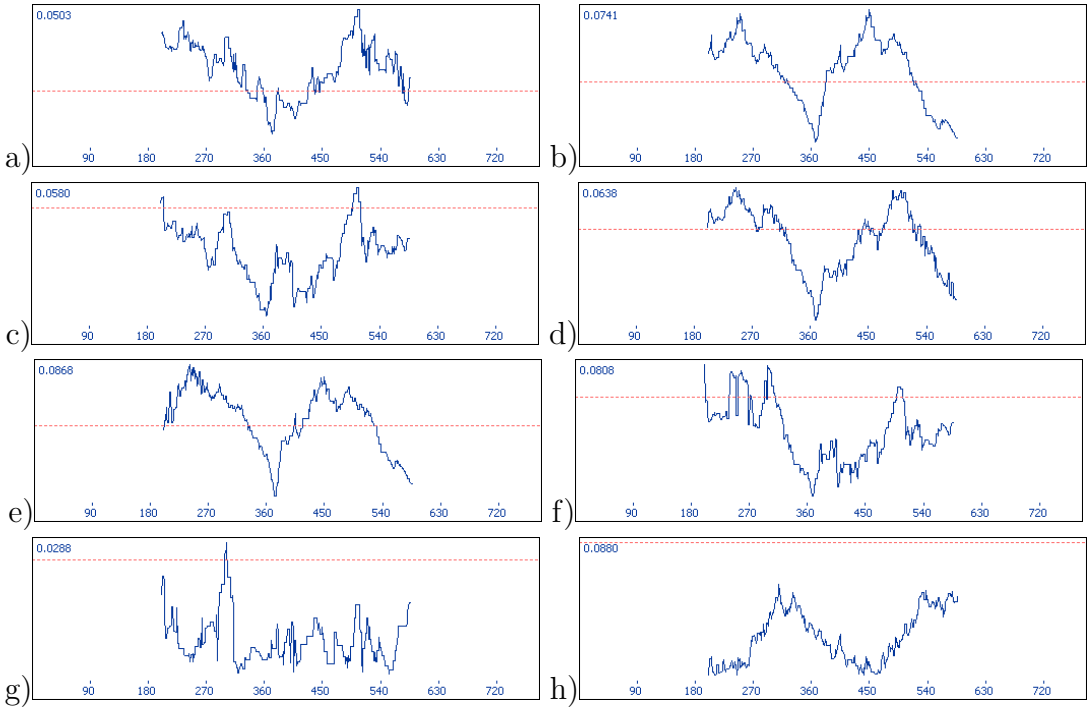


Figure 5.4: Selected subset of argF alignments analysed with DSS

Presented in the subfigures are the results of using DSS with a window size of 400 base pairs on the following alignments: (M,Mu,C,P), (M,Mu,L,G), (M,Mu,P,G), (M,C,L,P), (M,C,L,G), (M,C,P,G), (M,L,P,G), and (F,Mu,L,G). These alignments were chosen according to the results presented in figure 5.2, using phylogenetic networks. The horizontal axis represents the sites along the sequence alignment and the vertical axis the DSS statistic value. The dotted red line corresponds to the 95 percentile for the null hypothesis that there is no recombination. Sustained peaks above this red line have strong indications of recombination at these sites. Subfigures a, b, d and e show the greatest indication for heterogeneous regions (recombination) along the alignments.

5.3 BARCE

The method, BARCE (phylogenetic HMM or a pHMM described in Husmeier and McGuire (2002)), is an implementation of a Bayesian model using MCMC to infer topology changes along DNA sequence alignments. This model is a predecessor to the PFHMM of Husmeier (2005) which is a predecessor to the improved PFHMM developed in this thesis (BARCE is similar to the model of Husmeier (2005) without the modelling of the ratefactors). The BARCE model simulations return the posterior probability of each phylogenetic tree topology at each of the sites in the sequence alignment. Subsection 1.9.1 describes the approach of using HMMs for detecting recombination along alignments which is the theoretical foundation that applies to BARCE. For the purposes of this chapter, the statesequence of the topologies inferred by the model, $\mathbf{S}$, is the result of interest.

5.3.1 Definition

The topology state sequence $\mathbf{S} = (S_1, \ldots, S_N)$ is defined as in eq 1.73. The branch length vector, $\mathbf{w}$, and the nucleotide substitution parameters, θ are included in the model as in eq 1.71. For the probability of the data at each site t; $P(\mathbf{y}_t|S_t, \mathbf{w}, \theta)$ (emission probability) is calculated using the Kimura nucleotide substitution model described in subsection 1.3.3. The joint distribution is given as in eq 1.87 where ν_S is replaced with ν. Rate heterogeneity is not modelled in BARCE, and the PFHMM (described in subsection 1.9.4) has an independent chain for fitting the ratefactors along the sites of the sequence alignment. The posterior probability of the state sequence is found by integrating out all the other parameters $P(\mathbf{S}|\mathcal{D}) = \int P(\mathbf{S}, \nu, \theta, \mathbf{w}|\mathcal{D}) d\nu d\mathbf{w} d\theta$. This integral is computed via an MCMC simulation in a Metropolis-Hastings and Gibbs-within-Gibbs scheme:

$$\mathbf{S}^{(i+1)} \sim P\left(\cdot|\nu^{(i)}, \mathbf{w}^{(i)}, \nu^{(i)}, \mathcal{D}\right) \tag{5.3}$$

$$\nu^{(i+1)} \sim P\left(\cdot|\mathbf{S}^{(i+1)}, \mathbf{w}^{(i)}, \theta^{(i)}, \mathcal{D}\right) \tag{5.4}$$

$$\mathbf{w}^{(i+1)} \sim P\left(\cdot|S^{(i+1)}, \nu^{(i+1)}, \theta^{(i)}, \mathcal{D}\right) \tag{5.5}$$

$$\theta^{(i+1)} \sim P\left(\cdot|S^{(i+1)}, \mathbf{w}^{(i+1)}, \nu^{(i+1)}, \mathcal{D}\right) \tag{5.6}$$

This inference scheme is similar to the scheme of the PFHMM as it is presented in subsection 2.12.3. The inference of the topology state sequences (eq 5.3) was done via the Gibbs-within-Gibbs scheme described in the section A.9 and presently it

is done via the stochastic forward-backward algorithm.

Applying BARCE is more computationally demanding than applying the DSS statistic but less than would be required to calculate the posterior probability for the topologies with the improved PFHMM. One of the benefits of an analysis with this model over DSS is that it does not rely on the distance measures which have the intrinsic failures mentioned. The uncertainty between the choices of topologies along the sites of the alignment is also captured with BARCE. Using this method requires that the MCMC simulations converge and this is checked by running simulations multiple times under different initial configurations and iteration limits to test whether equivalent results are produced.

5.3.2 Application of BARCE

The work performed with the DSS statistic, presented in figure 5.4, highlighted certain sequence alignments for further analysis. The alignments shown in subfigures a), b), d) and e) are used in the following analysis using BARCE. These subfigures correspond to the groups of strains (M,Mu,C,P), (M,Mu,L,G), (M,C,L,P) and (M,C,L,G). The full meaning of these abbreviations are given in the caption of figure 5.1. These alignments are selected because according to the DSS statistics there is strong support for the existence of break points (recombination events) in these alignments, shown in figure 5.4.

For the MCMC simulations 1.2M iterations were chosen for the burnin phase and then 2M sampling iterations were sampled every 100 iterations, resulting in a set of 20K points from which the posterior probability of the topologies is calculated. The sampling rate is the same as the default of the BARCE TOPALi application. The recombination break points can be inferred from the posterior distribution of the topologies along the sites (**S**). The posterior distribution of the topologies has a choice between 3 topologies for the alignments of 4 sequences. The convergence for the simulations was checked by running independent simulations and similarity was used as an indication of convergence. The default number of iterations used by TOPALi for the burnin is less than the 1.2M used in this work. The adjustment was made to use 1.2M iterations for the burnin because after this point no further improvement in the stability of the converged results was observed. BARCE was used through TOPALi (described in 5.2.2) to run these simulations.

Figure 5.5 shows the sets of 3 plots for the mean posterior distributions of the topologies at each site on the alignment, sampled as in eq 5.3. In each sequence alignment where BARCE is run, 3 plots are produced showing posterior probabilities of the hidden states **S**. The 3 plots represent the topologies grouping first and second strains together, the first and third, and then the first and fourth strains. These topology plots are placed clockwise in the figure. Subfigures a) to c) represent the 3 topologies of the sequence group (M,Mu,C,P), subfigures d) to f) the topologies for group (M,Mu,L,G), g) to i) group (M,C,L,P), and j) to l) (M,C,L,G). Which letters represent which strains is stated in section 5.1 and in the caption of figure 5.1.

From the results the second and forth alignments show the strongest signals for topology break points due to recombination events. The first and third alignments have less consistent support for any particular topology along the sites. In the first and third alignments the mosaic structure of the topologies is not as easy to distinguish as in the results of the second and forth alignments. An explanation for this can be given by looking at the phylogenetic network of figure 5.1. The first alignment (subfigures a-c) contains sequences C and Mu which are closely related to each other with a relatively small trapezoid area between them. There is less information to distinguish between these two sequences. With the third alignment, it is possible that a recombination event groups both strains L and P creating an ambiguous signal. When not grouped together the small difference in 2 remaining topologies creates an oscillating signal. If there was more data the differences in the posterior would grow making a clearer inference possible. The same reasoning goes for the third alignment.

5.4 Application to Neisseria alignments chosen in literature

This section uses the sequence alignments of 4 strains as data with the DSS statistic, BARCE and the improved PFHMM. The alignments used are the pruned subset of the *Neisseria* data set in Zhou and Spratt (1992).

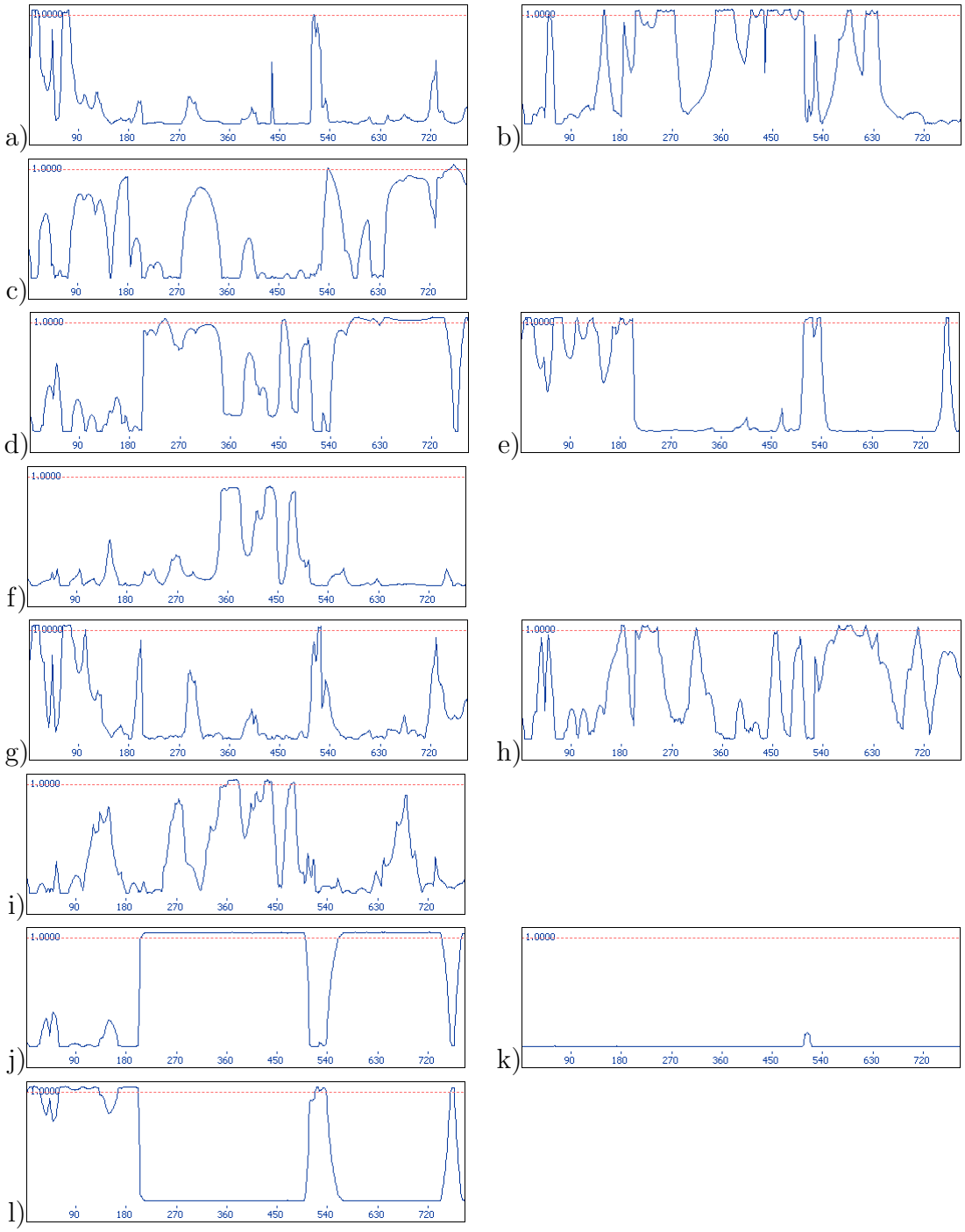


Figure 5.5: 4 *Neisseria* alignments run with BARCE

The choice of alignments is made according to the preliminary DSS analysis presented in figure 5.4. The vertical axis is the posterior probability of a particular topology at each site indexed on the horizontal axis. The dashed line is the 95 percent posterior probability. Subfigures a to c show the posterior distribution of the 3 topologies for the alignment of strains (M,Mu,C,P), subfigures d to f the group (M,Mu,L,G), g to i group (M,C,L,P), and j to l group (M,C,L,G). The 4 groups of subfigures presented have 3 plots for the different tree topologies. The second and forth alignments show the strongest indications of recombination with break points around sites 200 and 540.

5.4.1 Application of the DSS statistic and BARCE to Neisseria alignments chosen in literature

This subsection presents the results of using BARCE and DSS on the alignments of 4 sequences that were used in Zhou and Spratt (1992) and Husmeier (2005). The work of this thesis is closely based on that of Husmeier (2005) whose model is extended here. Examining the same *Neisseria* alignments used in these papers provides confirmation that this work is performing correctly, since all the strains discussed in the significant prior work are also present in the simulations in this work.

The paper of Zhou and Spratt (1992) uses the strains x64860 of N. gonorrhoeae, x64861 N. meningitidis, x64866 N. meningitidis, and x64869 of N. cinerea. Following the naming convention in this chapter we give them an abbreviated representation G, M_1, M_2 and C for x64860 of N. gonorrhoeae, x64861 N. meningitidis, x64866 N. meningitidis, and x64869 of N. cinerea respectively.

The results of the DSS and BARCE simulations for the (G,M_1,M_2,C) alignment are shown in figure 5.6. Subfigure a) shows the DSS result which can be compared to the previous results of figure 5.4. The horizontal axis represents the sites in the alignment and the vertical axis the DSS statistic. The dashed red line is the 95 percent threshold for the null hypothesis of there being no recombination. A window of 400 base pairs is used. Subfigures b) and c) show the BARCE results where the horizontal axis indicates the sites in the alignment and the vertical axis is the posterior probability of the particular topology at that site. Subfigure b) groups together strains M and G together as in subfigure l) of figure 5.5, and subfigure c) groups together both strains of meningitidis together. The third topology is excluded as it contributes a negligible posterior probability. The DSS and BARCE results both show a recombination event in the same region of the alignment; approximately at site 180. For the simulations with BARCE 40K samples were returned, and 1.2M iterations for the burnin stage was given. The simulations were consistent over multiple runs, indicating the stability of the result.

The *Neisseria* sequences used in Husmeier (2005) used strains N. gonorrhoeae X64860, N. meningitidis X64866, N. cinerea X64869, and N. mucosa X64873 (abbreviated as G, M, C and Mu respectively). In the list of sequences analysed with phylogenetic networks this alignment corresponds to subfigure 13) in figure 5.2.

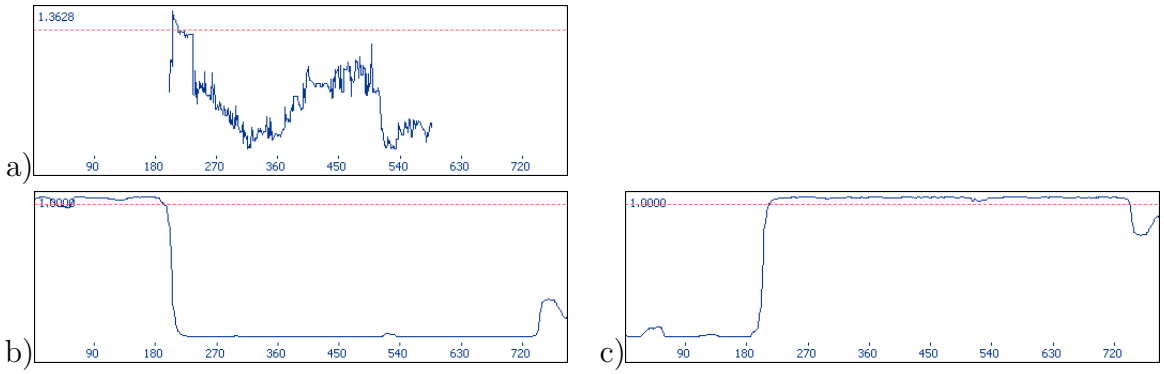


Figure 5.6: The results of using TOPAL and BARCE on the alignment of Zhou and Spratt (1992)

The alignment of Zhou and Spratt (1992) with strains G, M_1, M_2, and C is analysed with the DSS statistic and BARCE. Subfigure a) shows the DSS analysis performed with a 400 base pair window. The horizontal axis shows the sites in the alignment, the vertical axis is the DSS value, and the dashed red line is the 95 percent confidence threshold for no recombination under the null hypothesis. A DSS peak can be seen around site 200. Subfigures b) and c) show the result of using BARCE for the topology grouping strains G with M_1, and M_1 with M_2 respectively. The remaining topology grouping M_1 with C contained less than 5 percent of the total probability along the sites. A change in the majority of the posterior probability can be seen around the same region as that indicated at around site 200.

Figure 5.7 shows the results of the DSS and BARCE analysis on these selected sequences (as is done for the alignment of Zhou and Spratt (1992) shown in figure 5.6). Subfigure a) shows the result of using the DSS statistic. As before the horizontal axis represents the sites in the alignment and the vertical axis the DSS statistic. The dashed red line is the 95 percent threshold for the null hypothesis of there being no recombination. There is a long region where the DSS statistic is above the 95 percent threshold. This region begins approximately at the same site as the previous alignment in figure 5.6, site 180. There appears to be a recombination point centered around the site 540 which is not included in the Zhou and Spratt (1992) alignment. A window of 400 base pairs was used with DSS. Subfigures b, c and d group the strains G with M, G with C, and M with Mu respectively. The recombination point seen between figures b) and d) is also seen from the DSS result in the subfigure a) here and with the results of figure 5.6. The peak around site 540 found with DSS can also be seen in the change of posterior probabilities between the topologies grouping M with G (in subfigure b) and M with Mu (in subfigure d). This inferred recombination point was not present in the analysis done with the alignment of Zhou and Spratt (1992). For the simulations run with BARCE an 80K sample size was used (with 100 iteration interval between each sample taken), and 1.2M iterations for the burnin stage which was consistent over multiple runs.

The biological conclusions from these results are discussed fully in the conclusion section of this chapter, section 5.5. From a methodological perspective it is re-assuring that new topology break points were not found in these alignments in comparison to the previous analysis. The beakpoint at site 540, seen with BARCE, is not strong enough with the DSS results shown here, but using different sequence alignments with DSS does show the 540 breakpoint region (figure 5.6). The strains of these alignments also overlap with the main choices given before. These results using DSS and BARCE draw the same mosaic structure of the topologies as the papers Zhou and Spratt (1992) and Husmeier (2005) which used different methods.

5.4.2 Application of the improved PFHMM

The phylogenetic factorial hidden Markov model (PFHMM) is introduced in subsection 1.9.4. The improved PFHMM that is used here is presented in chap-

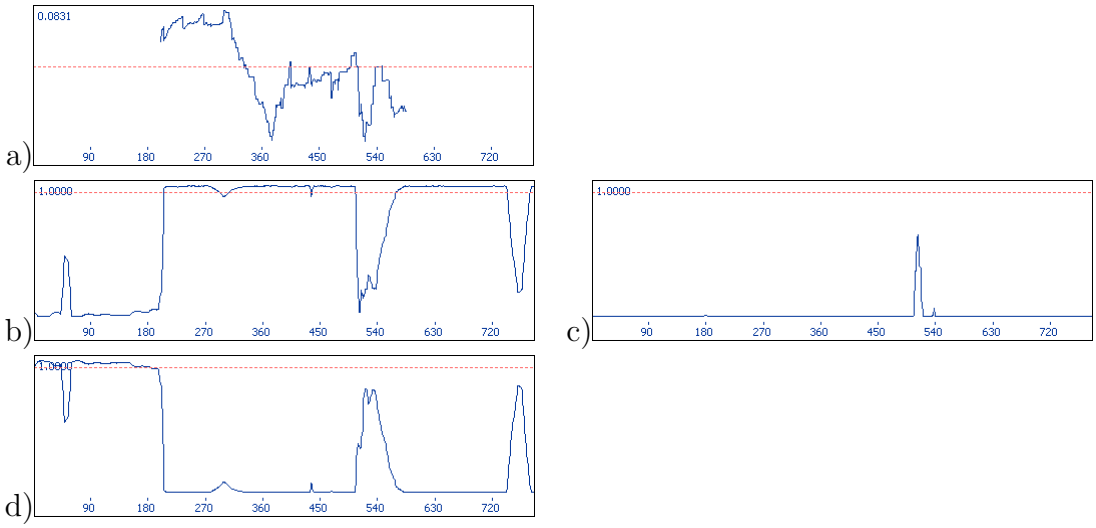


Figure 5.7: The DSS and BARCE analysis of the Neisseria alignment used in Husmeier (2005)

The *Neisseria* sequences used in Husmeier (2005) G, M, C and Mu are analysed for recombination break points using the DSS statistic and BARCE. Subfigure a) presents the results of using the DSS statistic with a window of 400 base pairs. The horizontal axis shows the sites in the alignment, the vertical axis is the DSS value, and the dashed red line is the 95 percent confidence threshold for no recombination under the null hypothesis. A DSS peak can be seen around site 200 and 540. The BARCE results for the posterior probabilities of each topology at each site is shown in subfigures b), c), and d). These figures show the topologies grouping the strains G with M, G with C, and M with Mu respectively. The recombination events predicted with DSS can be seen in the BARCE results having changes in the posterior around sites 200 and 540.

ter 4. There are 2 alignments which are chosen to be analysed with the improved PFHMM. The first is the alignment containing the sequences (M,Mu,L,G), and the second is the (M,C,L,G) alignment. The abbreviations of the strains are as explained in the caption of figure 5.1. The reason for choosing these two alignments is because they showed strong signals of recombination when processed using DSS and BARCE. From the subsection 5.4.1 the alignments (G,M_1,M_2,C) and (G,M,C,Mu) were used to compare the alignments used in the published literature, and from them a clear signal of recombination is produced.

The strains which appear in alignments showing the clearest topology mosaic structures are (M,Mu,C,L,G), and two alignments are created from this set. The strains M and G have been essential in all of the alignments which have strong indications for recombination, and are chosen to be present in both alignments. Examining again the phylogenetic network of figure 5.1, there is a relatively smaller distance between Mu and C than between other pairs of strains. Two alignments are made by having either Mu or C present.

The results of running the improved PFHMM with the alignment containing the strains (M,Mu,L,G) is shown in figure 5.8. There were 450 burnin steps and 450 sampling steps for the Gibbs sampling scheme. For the sampling of the ratefactors, branch lengths and the vector of the relative codon substitution rate; 300 burnin steps and 600 sampling steps were given. The PSRF factor (Gelman and Rubin (1992)) was below 1.3 for the sampling phase for the parameters of the simulation which includes the branch lengths and ratefactors. The results show an intricate mosaic structure of the topologies in subfigure a). The 3 subplots are for the 3 topologies grouping M with Mu, M with L and M with G from the top to the bottom respectively. The horizontal axis are the sites in the alignment and the vertical axis the posterior probability of the topology at a site. In subfigure b) the mean posterior probability of the ratefactors is plotted according to eq 4.48 along the sites. The 66 and 95 percent credibility intervals are plotted as a dashed and a dotted line about the mean which are hard to identify as they lay close to the mean in this simulation. Subfigure c) shows the histogram of the number of ratefactors allocated from the RJMCMC scheme. The mosaic structure in the topologies and ratefactors is intricate. The break points for the topologies at site 200 can be seen as expected from the figures of subsection 5.3.2 and subsection 5.4.1. There is another topology break point before site 100 indicating that the recombinant region is smaller than inferred previously. The sharp peak at site 540, for the

topologies found in figure 5.7 and can also be seen in these results. Roughly between the sites 300 to 500 there is a topology switch to group strains M with G which is discussed in the caption of figure 5.9 and the text accompanying the figure. When examining subfigure b) the positions for the break points around site 100 and the 300-500 base pairs region are visible here as well. There are 2 ratestates (from subfigure c) of approximate values 0.5 and 1.

The results of using the improved PFHMM with the (M,Mu,L,G) alignment will now be compared to the results of using DSS and BARCE with the same alignment. From figure 5.8 the subfigures a) and b) are presented again as the subfigures a) and b) in figure 5.9. Subfigures c), d), and e) show the BARCE result for the 3 topologies grouping M with Mu, M with L and M with G. The horizontal axis represents the sites in the alignment and the vertical axis the posterior probability of the topology at a given site. The DSS analysis is shown in subfigure f) with the horizontal axis being the sites in the alignment and the DSS value the vertical axis. The dashed red line is the 95 percent threshold for the null hypothesis of there being no recombination. The results show that the improved PFHMM infers a recombination break point in the region of 0-100 which BARCE and DSS do not detect. The break point at the region 180 to 200 is detected by each method, as is the abrupt topology change around site 500. The break points for the region 300-500 indicate a grouping of strains M with G, which is unexpected as it is not present in the results of other authors. If the simulations had not converged then this may manifest as a change in the ratefactor values shown in b) which for this region could create a spurious topology change. To test whether this was the case the PSRF was computed which showed convergence had occurred, additionally other alignments containing either one of these strains singly did not show these artifacts. Referring again to the phylogenetic network reconstruction, strain G would have a favourable grouping with strain M in the absence of non-linear substitutions (eg. recombination) which is a possible scenario. The alignment that differs by removing Mu and replacing with it with C shown later in figure 5.10, and that does not exhibit this effect in the region, showing that C contains a close relationship to M in this region.

The results of running the PFHMM with alignment (M,C,L,G) is shown in figure 5.10. This alignment was analysed with BARCE in the figure 5.5 (subfigures j, k, and l). For the simulation 450 Gibbs burnin steps and 450 Gibbs sampling steps were given. The sampling of the ratefactors, the branch lengths and the

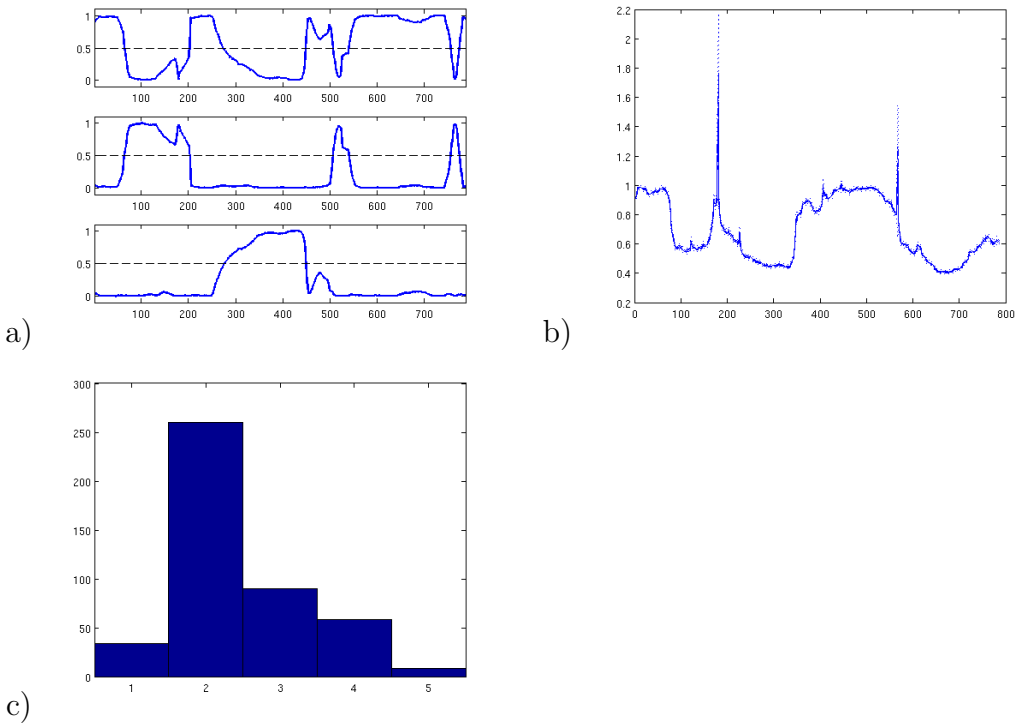


Figure 5.8: The alignment of (M,Mu,L,G) analysed with the improved PFHMM

The results of running the improved PFHMM with the alignment containing the strains (M,Mu,L,G). The posterior probabilities of the topologies along the sites is shown in subfigure a). The 3 subplots are for the 3 topologies grouping M with Mu, M with L and M with G from the top to the bottom respectively. In subfigure b) the mean posterior probability of the ratefactors is plotted according to eq 4.48 along the vertical axis and the horizontal axis are the sites along the alignment. The 66 and 95 percent credibility intervals are plotted as a dashed and dotted line about the mean. Subfigure c) shows the histogram of the number of ratefactors allocated from the RJMCMC scheme. Topology break points can be seen around sites 100, 200, 300, 500, and 540. There are 2 ratefactors which take the values approximately of 0.5 and 1. The ratefactor break points coincide with the topology break points around sites 0-100 and 300-500.

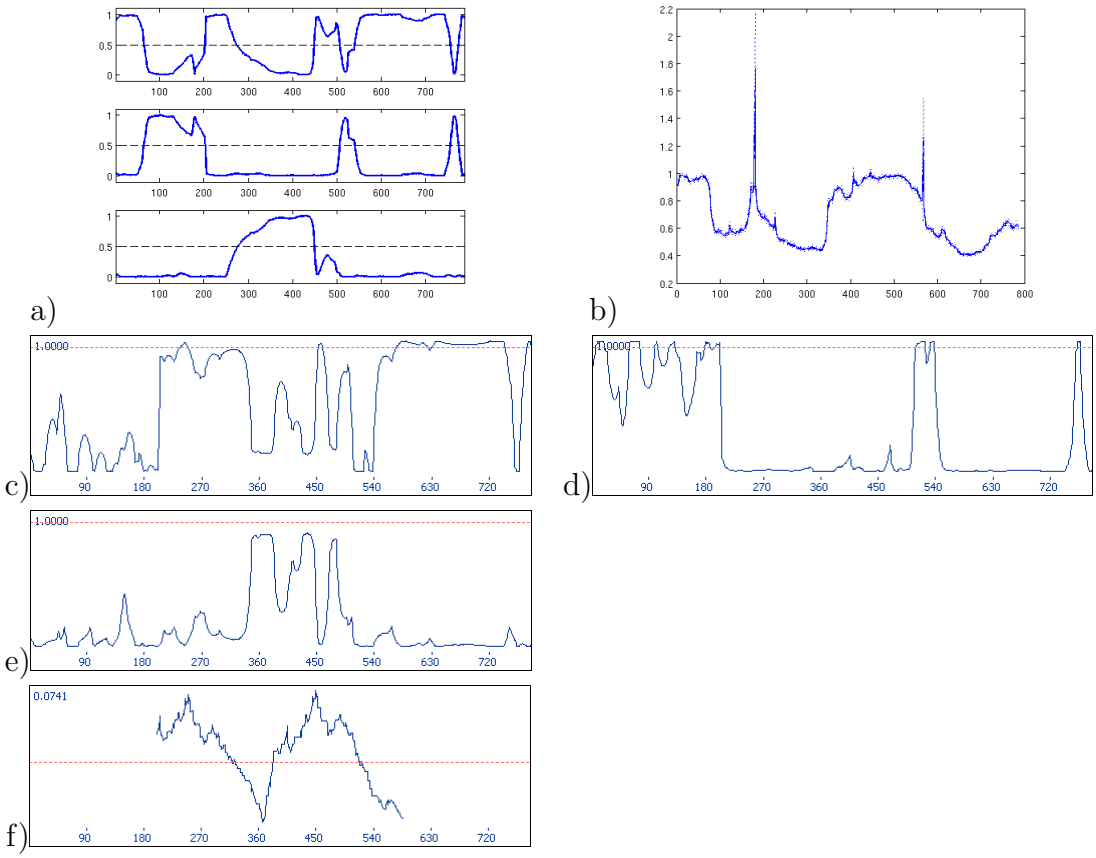


Figure 5.9: Layout of the analysis on group (M,Mu,L,G) with the improved PFHMM, BARCE and TOPAL

The results of using the PFHMM with the (M,Mu,L,G) alignment, shown in subfigures a) and b), is compared with the BARCE results in c), d) and e) and with DSS in f). Subfigures c), d), and e) show the topologies grouping M with Mu, M with L and M with G respectively. The horizontal axis represents the sites in the alignment and the vertical axis the posterior probability of the topology at a given site. For subfigure f) the horizontal axis are the sites in the alignment and the DSS value the vertical axis. The dashed red line is the 95 percent threshold for the null hypothesis of there being no recombination. From a), a recombination break point in the region of sites 0-100 can be seen which BARCE and DSS do not detect, but is seen in b) for the change of value of the ratefactors. The break point at the region 180 to 200 is picked up in each method. The abrupt topology change around site 500 is found by the group of methods as well. For the region 300-500 the PFHMM groups strains M with G, and a change in the same region is seen for the ratefactors in b).

values for the relative codon vector had 300 burn-in steps and 600 sampling steps. The PSRF factor was below 1.3 for the sampling of the parameters during the simulation. Subfigure a) of this same figure shows a set of 3 plots for the posterior probability of each topology along the sites of the alignment. From the top to bottom the plots are of the groupings of the strains M with C, M with L, and M with G. The horizontal axis are the sites in the alignment and the vertical axis the posterior probability. In subfigure b) the mean posterior probability of the ratefactors is plotted according to eq 4.48 along the sites. The 66 and 95 percent credibility intervals are plotted as a dashed and a dotted line about the mean which may be hard to identify as they lay close to the mean. Subfigure c) shows the histogram of the number of ratefactors allocated from the RJMCMC scheme. The topologies show a recombination event around site 200 and a peak approximately at site 540 which is very close to the predicted results that were given by BARCE. The ratefactors along the alignment share a similar structure to the simulation results shown in figure 5.8 for strains (M,Mu,L,G) in the regions 0-100 and 300-500 where there are change points. The values that the ratefactors appear to switch between are approximately 0.5 and 1.0, and are similar to those in figure 5.8.

The comparison of the improved PFHMM results along with the DSS and BARCE results for the same alignment containing strains (M,C,L,G), is shown in figure 5.11. From figure 5.10 subfigures a) and b) are presented as subfigures a) and b) as well. Subfigures c), d) and e) show the results of the BARCE simulations with the topologies grouping M with C, M with L, and M with G respectively. The vertical axis is the posterior probability of the topology at each site of the alignment on the horizontal axis. Subfigure f) shows the result of using DSS on the alignment. The vertical axis is the DSS statistic and the horizontal axis the sites in the alignment. The red dashed line corresponds to the 95 percent confidence in the null model for there not being recombination. The improved PFHMM, BARCE and DSS infer a recombination event in the region of the sites 180-200. The narrow recombination region around site 540 is also found. This artifact at site 540 is inferred as a change in the ratefactor value in Husmeier (2005). When using the improved PFHMM a different mosaic structure for the ratefactors is found than in Husmeier (2005). Around this site a topology change is present as well, as with the alignment (M,Mu,L,G). Here the ratefactors show a different mosaic structure which was not the case with the previous alignment.

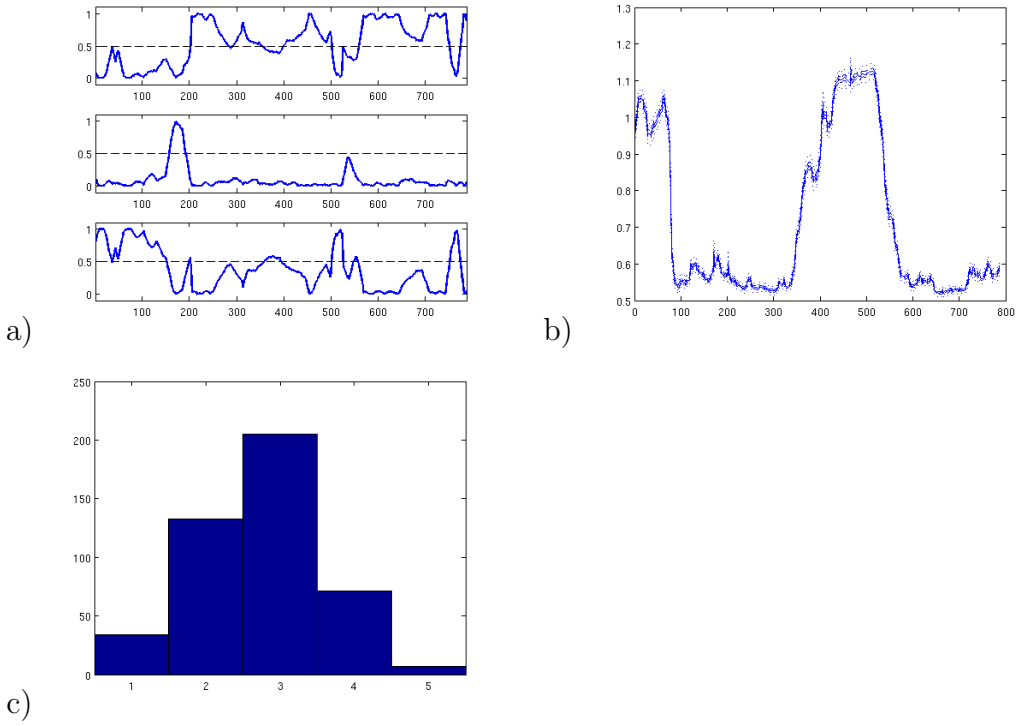


Figure 5.10: Results of using the improved PFHMM with the alignment (M,C,L,G) Subfigure a) shows a set of 3 plots for the posterior probability of each topology along the sites of the alignment. The top to bottom of the plots groups the strains M with C, M with L, and M with G. The horizontal axis represents the sites in the alignment and the vertical axis the posterior probability of each topology. In subfigure b) the mean posterior probability of the ratefactors is plotted according to eq 4.48 along the sites. The 66 and 95 percent credibility intervals are plotted as a dashed and a dotted line about the mean. Subfigure c) shows the histogram of the number of ratefactors allocated from the RJMCMC scheme. The topologies show a recombination event around site 200 and there is a spike topology change approximately at site 540 which is very close to the predicted results that were given by BARCE. The ratefactors have change points in the regions 0-100 and 300-500, and appears to switch between are approximately 0.5 and 1.0.

5.5 Conclusions

This chapter provides evidence for the presence of recombination and rate heterogeneity in strains of *Neisseria*. The set of sequences were taken from the work of Zhou and Spratt (1992) who submitted a data set of *Neisseria* strains and conducted a study to detect recombination amongst them. The improved PFHMM, described in chapter 4, is applied to investigate the published recombination events of both Zhou and Spratt (1992) and Husmeier (2005) and tests whether new events can be found. To reduce the computational burden, before applying the improved PFHMM to the real data set a series of methodologies are applied in stages to find the best candidate alignments of 4 sequences. The biological and methodological results will be discussed for each methodological stage in the order that they appear in this chapter.

The improved PFHMM can handle the computational requirements for analysing sequence alignments of 4 strains within reasonable time constraints. As discussed earlier, the number of topologies grows at a super-exponential rate in the number of strains, so unfortunately calculating results for 5 strains is beyond the scope of this work. The set of *Neisseria* sequences from Zhou and Spratt (1992) contains more than 4 strains, and therefore requires that a selection be made. Phylogenetic Networks as described in section 5.1 are applied to the complete set of Neisseria sequences as it is much less computationally demanding and gives a very coarse analysis which is suitable for an initial simplification step. The results of using phylogenetic networks are shown in figure 5.1 and the labels on the leaves of the network are an abbreviation of the names used in the EMBL database. The network produced provides strong evidence for the simplification of the set of strains by using only a single N.*meningitidis* strain to represent the set of closely placed meningitidis strains. The network surrounding this family of strains presents little indication of possible recombination events or other non-linear evolutionary events between the family of meningitidis strains. In their investigation Zhou and Spratt (1992) applied their approach to different meningitidis strains, but the same features can be found using only a single meningitidis strain as shown in Husmeier (2005) and the analysis of this chapter in figure 5.6. The split areas (rectangular areas) created by phylogenetic networks were largest

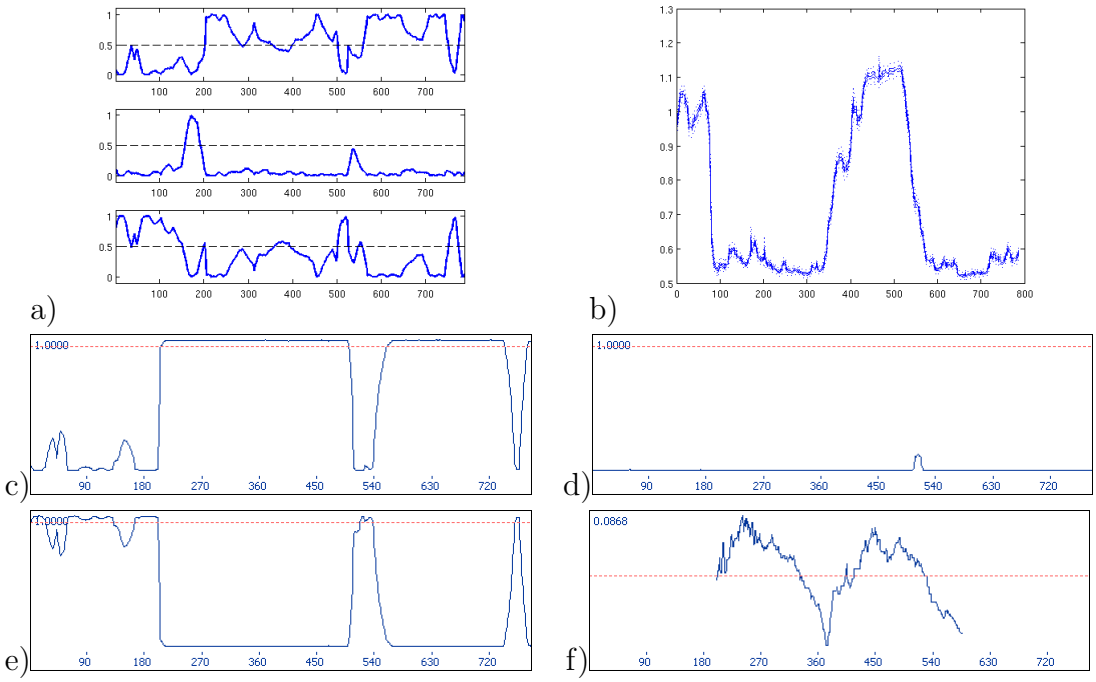


Figure 5.11: Comparison of the analysis of the alignment (M,C,L,G) with the improved PFHMM, BARCE and DSS

Subfigures a) and b) are copied to here from figure 5.10. Subfigures c), d) and e) show the results of the BARCE simulations with the topologies grouping M with C, M with L and M with G respectively. The vertical axis is the posterior probability of the topology at each site of the alignment shown on the horizontal axis. Subfigure f) shows the result of using DSS on the alignment. The vertical axis is the DSS statistic and the horizontal axis the sites in the alignment. The red dashed line corresponds to the 95 percent confidence in the null model for there not being recombination. The improved PFHMM, BARCE and TOPAL infer a recombination event in the region of the sites 180-200, and a sharp recombination event around site 540 is also found. The mosaic structure of the topologies is different to that of the ratefactors.

for the strains of gonorrhoeae (G), lactamica (L), and polysaccharea (P). In the analysis that followed, it was found that the inclusion at least 2 of these 3 strains are essential in order to be able to observe strong indications for recombination and rate heterogeneity. Producing a phylogenetic network therefore assisted in a significant simplification of the dataset and reduced the size of the exploration for the subsequent analysis. This simplification comes with strong evidence that there is a linear evolutionary relationship between strains of meningitidis. An insight for which of the strains would be important for detecting recombination was also found by using this method.

Subsection 5.1.2 constructs phylogenetic networks from the set of possible alignments of 4 sequences and displays them in figure 5.2. For alignments of only a few sequences, the magnitude of the conflicting signal for a single phylogenetic tree construction is easy to evaluate with this method as it is simply the area of the trapezoid associated with the split. The caption lists the networks displaying the largest conflicting signals (splits) by the size of the trapezoidal area they contain. The selection of alignments to be further analysed is based on the area of the trapezoidal region, with larger areas indicating more recombination possibilities. This selection frequently includes strains G, M, and C which are common between the alignments chosen in Zhou and Spratt (1992) and Husmeier (2005). The overlap on the choice of strains supports the earlier conclusions based on phylogenetic networks giving results consistent with the published literature.

After the analysis using phylogenetic networks, the DSS statistic (described in section 5.2) is then used. Following a similar motivation to the initial filtering approach using phylogenetic networks, DSS is computationally not as demanding as the more complex methods subsequently applied and produces a more detailed analysis for possible recombination than basic phylogenetic networks. This makes it suitable for use as a computationally efficient second stage in the improvement of the analysis. DSS provides a measure for comparing the possibility of topology break points along the sites of an alignment which phylogenetic networks can not provide and DSS also has a stronger statistical foundation for detecting recombination. DSS is applied to the selection of alignments noted in the caption of figure 5.2. Two extra alignments of 4 sequences are included to ensure that there is enough diversity in the sequences. The results are presented in figure 5.4 showing that there are 4 alignments (subfigures a, b, d, and e) which have long regions of sites with DSS values above the 95 percent confidence interval for the null model

of no recombination. These subfigures correspond to the alignments (M,Mu,C,P), (M,Mu,L,G), (M,C,L,P), and (M,C,L,G). All 4 of the selected DSS value plots show a similar pattern in the placement of peaks above the 95 percent confidence interval. Amongst these 4 selected alignments the DSS statistic surpasses the 95 percent confidence interval (dashed horizontal line) approximately at the sites 180-300 and 450-550. The comparison of the plots where strain M (for meningitidis) is present or not shows that strain M is a requirement for observing evidence of recombination. Also, so are either C or Mu (cinerea or mucosa), and then either G/L/P (gonorrhoeae/lactamica/polysaccharea) needed in the alignments for evidence of recombination to be visible using these methods. A reasonable interpretation (also bearing in mind the phylogenetic network from figure5.1) is that between meningitidis and cinerea/mucosa there are no (or few/short) recombination events, and that from the ancestor of gonorrhoeae/lactamica/polysaccharea there is at least one recombination event.

The method BARCE is introduced in section 5.3, and is used to continue the refinement process by applying the method to the alignments selected by the DSS analysis. BARCE uses an HMM to model the topologies as the hidden states and is the same as the model of Husmeier (2005) but BARCE does not model the rate heterogeneity along the sequence alignment. DSS is computed from pairwise distances between sequences and because of this suffers from intrinsic information loss. The model of BARCE builds a phylogenetic tree and estimates the uncertainty in the space of topology choices. This allows the model to assess candidate topologies and returns the posterior probability of the topologies along the sites. The inference scheme for this more complicated model is more computationally demanding and simulations need to be run multiple times to check for convergence.

The results of using BARCE on the alignments (M,Mu,C,P), (M,Mu,L,G),(M,C,L,P), and (M,C,L,G) is shown in figure 5.5. Although the simulations are run multiple times and for an extended number of iterations, there are significant oscillations in the posterior probabilities for the first and third alignments. The fourth alignment does not exhibit these oscillations and the second alignment does but to a lesser extent, thus these alignments are preferred. Both the second and fourth alignments exhibit topology changes (plausible recombination events) in the same regions indicated by the DSS analysis; in the region of sites 180-300 and 450-500. In further BARCE simula-

tions for the alignments of Zhou and Spratt (1992) and Husmeier (2005) these oscillations were not present. Examination of the strains used in these studies suggests the tentative conclusion that oscillations in the topologies may be due to similar likelihoods when specific strains are included in an alignment, and causes the topology state transition parameter to have a value too low. Considering (M,Mu,C,P) it can be seen that the differences between Mu and C are small. And for the alignment (M,C,L,P) the distances between these strains may not be large enough in respect to M. This would explain why the oscillations are not reported in Zhou and Spratt (1992) and Husmeier (2005).

In subsection 5.4.1 BARCE and DSS are applied to the sequence alignments of Zhou and Spratt (1992) and Husmeier (2005). For Zhou and Spratt (1992) the strains of the alignment are (G,M_1,M_2,C) corresponding to gonorrhoeae, meningitidis (1 and 2), and cinerea with the results shown in figure 5.6. For Husmeier (2005) the alignment uses the strains (G,M,C,Mu) which correspond to gonorrhoeae, meningitidis, cinerea, and mucosa with results shown in figure 5.7. The DSS and BARCE analysis of the alignment (G,M_1,M_2,C) shows a single possible recombination event within sites 180-200. This region is consistent with the first of the two suspected recombination events found in previous results of figure 5.5. The equivalent analysis is repeated for the alignments of Husmeier (2005) and shows that both DSS and BARCE predict 2 recombination regions which are analogous to those independently found in figure 5.5. This supports the previous hypothesis that meningitidis and gonorrhoeae are a requirement for observing the recombination event in the 180-200 site region.

Section 5.4.2 applies the improved PFHMM which is discussed in chapter 4. There is an increased computational demand that comes with the complexity of the model. Simulations are run with 2 different sequence alignments. The choice of which alignments of 4 sequences to apply is made on the basis of the previous results shown in figure 5.5. The first alignment used is (M,Mu,L,G) whose strain names are meningitidis, mucosa, lactamica, and gonorrhoeae respectively. The PFHMM results for this alignment are shown in figure 5.8. The topology changes anticipated after the BARCE simulations, are present at the approximate sites 180-200 and 540 shown in figure 5.8. There are new observed topology changes as well which indicate that in the evolutionary history of meningitidis there are other possible recombination events. These are in the site regions of 0-100 and 300-450 which also correspond to break points in the ratefactors. These new

observations could be due to a lack of convergence of the simulation or a lack of sufficient data to distinguish between events. In considering this scenario the results for the same alignment run with the DSS statistic and BARCE is shown for comparison in figure 5.9 where the 0-100 recombinant region is not observed but the 300-500 region is. These topology changes are observed by both BARCE and improved PFHMM. As BARCE does not account for rate heterogeneity whilst it is possible to conclude from the BARCE findings that these topology changes may be due to either rate heterogeneity changes or recombination events, the improved PFHMM observations would not support this conclusion and instead suggests that the topology changes are in fact due to recombination events.

The improved PFHMM is also applied to the alignment of (M,C,L,G) with strain names meningitidis, cinerea, lactemica, and gonorrhoeae. The results are shown in figure 5.10 and the comparison with the use of the DSS statistic and BARCE is shown in figure 5.11. The structure of the ratefactors mirror the plot for the (M,Mu,L,G) alignment with the approximate same values for the 2 ratestates and break point positions. The anticipated topology break points are present (site regions 180-200 and 540) but not the new break points found with (M,Mu,L,G) which are only very weakly supported. This draws possible conclusions about the impact of the choice of strains on the results of both BARCE and the improved PFHMM. Husmeier (2005) inferred that the event at site 540 is due to ratefactor change being interpreted as a topology change. It should be noted that in making comparisons between the results of Husmeier (2005) and the findings presented here that this work uses slightly different alignment sequences (not all strains are identical in the alignment) and for these sequences the findings suggest that this is in fact a topology change.

Due to the small size of the alignments and the number of features which may be contained, it appears that it is the presence of a vague posterior rather than the methodological developments that is the cause of the limits for drawing further conclusions. A local optima spreading over a large region of the parameter space could also create ambiguous results. Another possible scenario is that the model does not capture additional evolutionary developments contained in these alignments. A possible solution would be to test the method with a more informative prior on the topology state transition parameter. The improved PFHMM findings putatively suggest the existence of a recombinant region in the sites 300-500 and also a new ratefactor structure. For the findings reported here, the degree

of uncertainty is too high to confirm these observations, and so they remain as untested hypotheses.

Given the previous discussion in this chapter and the results of the simulations using the improved PFHMM, some final remarks about the methodology and biological aspects of the dataset can be made. The results of the improved PFHMM were well supported by the prior findings but also showed that on certain small sequences the model did not produce unambiguous results for the topologies. This failure was unexpected as the synthetic test cases used for development had not had the same properties as this particular real world DNA sequence alignment. The cause of the failure in this instance is believed to be due to the vague posterior caused simply by lack of sufficient data. There is less uncertainty in the results for the ratefactors which indicate a new structure of rate heterogeneity (which can be compared to the results in Husmeier (2005)). Also reassuring is that the number of sampled ratefactors is unimodal and has low variance showing that the model explored various numbers of components and stably focused on the most optimal number.

Use of the PFHMM appears to require more information between break points than was contained in some of the sequence alignments used. What may be concluded is that the recombination event inferred in literature between strains meningitidis and gonorrhoeae is also supported by the results found here. We also find that a more complex structure of rate heterogeneity is also very likely. Clearly this work has made significant progress in incorporating recombination models and rate heterogeneity into topology inference for phylogenetic trees, but the area is rich with unmined possibilities; future work would involve studies with sequences of a greater number of sites and with relatively fewer recombination events or regions of rate heterogeneity.

Chapter 6

Conclusions

In my thesis, I have investigated a possible shortcoming of three recent Bayesian methods for detecting recombination in DNA sequence alignments: the multiple change-point (MCP) model of Suchard *et al.* (2003), the dual multiple change-point (DMCP) model of Minin *et al.* (2005), and the phylogenetic factorial hidden Markov model (PFHMM) of Husmeier (2005). All three models assume separate branch lengths for different sites, which allows the branch lengths to be integrated out analytically. This reduces the computational complexity of the Bayesian inference scheme, which can now be formulated in terms of posterior distributions of the tree topologies and the nucleotide substitution parameters only. This makes this approach, which was first introduced by Tuffley and Steel (1997) under the name "no-common-mechanism" model, quite popular, and it has been applied in more recent works; see Lehrach (2008), Lehrach and Husmeier (2009), and Webb *et al.* (2009). The principle problem with the no-common-mechanism model is that the branch lengths are incidental rather than structural parameters. In my thesis, I have shown that a model with the no-common-mechanisms assumption is susceptible to the prediction of wrong tree topologies for certain branch length configurations (long branch attraction), and that it suffers from the same inconsistency (long-branch attraction) as the method of maximum parsimony. In particular, my study has shown that recombination detection methods using the no-common-mechanism model are susceptible to predicting spurious recombination events whenever branch-length configurations happen to fall near the boundary of the Felsenstein zone.

To address this difficulty, I have removed the site-independence assumption for the branch lengths. As a consequence, the analytic integration over the branch

lengths is no longer tractable, and they have to be sampled approximately from the posterior distribution with MCMC. To avoid an identifiability problem resulting from the fact that the global scaling of the branch lengths (defined by one of the two types of hidden states) is an additional independent parameter of the model, I have imposed a normalization constraint on the branch lengths. I have tested the proposed method on the same DNA sequence alignments as have been used in testing the other models, and found that it succeeded in avoiding the failure in the Felsenstein zone.

Further, I have explicitly modelled within-codon rate heterogeneity via a separate rate modification vector. In this way, the within-codon effect of rate heterogeneity is imposed on the model a priori, which facilitates the learning of the biologically more interesting effect of regional rate heterogeneity a posteriori. I have carried out simulations on synthetic DNA sequence alignments, which have borne out my conjecture. The previous model, which did not explicitly include the within-codon rate variation, has to model both effects with the same modelling mechanism. As expected, it was found to fail to disentangle these two effects. On the contrary, I have found that my improved model clearly separates within-codon from regional rate heterogeneity, resulting in more accurate predictions.

I have finally combined my models with the RJMCMC scheme of Lehrach and Husmeier (2009) so that the number of rate states is not defined a priori but sampled from the posterior distribution via RJMCMC. While the model of Lehrach and Husmeier (2009) is susceptible to long branch attraction and does not distinguish the within-codon effect of rate heterogeneity from long-range rate variation, the model proposed in my thesis deals with these effects. My simulations have shown that the model is capable of handling the combined complexity of the improvements made.

Appendix A

Appendix

A.1 Number of possible rooted and unrooted topologies for a given sequence alignment

Given a sequence alignment of m number of DNA sequences a range of possible candidate topologies exist to explain the ancestral relationship between the genetic data. These phylogenetic trees can be constructed as 'rooted' tree, or 'unrooted'. Rooted trees provide an ancestor for all of the taxa observed, and position it in the tree. Unrooted trees are uninformative about the position of the position of the root.

To determine the number of possible different rooted topologies we consider the phylogenetic tree built from m sequences, which will have m leaves. Moving up the tree from a leaf node the edges from the bifurcation coalesce to the common ancestor to the adjacent taxa is met. This reduces the number of edges by 1, and continues to reduce the number of edges by one for each step up the tree as the bifurcations coalesce. Therefore $m-1$ nodes as well as the m taxa leaves exist, and adding these two together gives $2m-1$ nodes in total ($2m-2$ edges since there is no edge above the root node that is considered).

For a phylogenetic tree that is unrooted $2m-2$ nodes exist, and $2m-3$ edges (branch lengths). This is because from the unrooted tree a root node can be added along any of the edges, increasing the number of nodes by 1 (as well as the number of edges). This then produces the same number of edges and nodes as for the rooted trees as explained in the previous paragraph. With there being $2m-3$ edges to choose from in adding the new node to be the root this will increase the

previous number of unrooted topologies by a factor of $(2m-3)$.

In determining the number of topologies possible for a sequence alignment of m taxa, we first consider unrooted trees produced. For $m=3$ there is only 1 possible unrooted topology, there are 3 edges, and adding another branch with a leaf node produces a tree with $m+1$ leaves. Since there were three edges originally the tree with $m+1=4$ leaves has 3 topology configurations having $2m-3=5$ edges. The tree with $m=4$ now has 5 edges where a branch and leaf node can be added, and for each of the 3 possible topologies there are 5 edges to add the new leaf giving a distinct topology as a result; so the number of topologies with $m=5$ is $3\times 5=15$. The number of edges is increasing by 2 for each extra leaf node added, when a new leaf node with branch is added the number of possible topologies that existed before are multiplied by the number edges it contains, and this is generalised by: $3\times 5\times 7\times\ldots\times(2m-5)=(2m-3)!!$. For rooted trees we consider that the same relationship holds as for unrooted trees but there 1 more node for the m sequences than with unrooted trees so in the series the next term is included with $(2m-3)!!$. Tree counting is discussed in detail in Felsenstein (1978b).

A.2 Branch lengths as the expected distance between sequences

The working for using the branch lengths to compute the distance between species in a phylogenetic tree via the branch lengths and the expected number of substitutions is shown. An example is made using the Jukes-Cantor model of nucleotide substitution described in subsection 1.3.2.

The vector of branch lengths for the phylogenetic tree is denoted with $\mathbf{w}$. The normalised branch lengths used later in the model is represented as $\dot{\mathbf{w}}$ because the ratefactors exist as well there to scale the normalised vector to different magnitudes without changing the direction of the vector. Subsection 1.5 defined the branch lengths in the nucleotide substitution model to be the product of the rate and the time; $w=(\rho)\times(t)$ (where t is for the time here and not positions/sites in the sequence alignment or in the HMM).

Some of the equations overlap with the section on the Jukes-Cantor model which are included here as well. The entries for the transition matrix (defined in

eq 1.3 and in eq 1.25) are

$$p_{i,j}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} & i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} & i \neq j \end{cases} \tag{A.1}$$

This is based on the Jukes-Cantor model defined in eq 1.21 where all non-identical character substitutions happen at the same rate, α. The rate ρ which is general can be substituted with the substitution rate α for the expression of a branch length $w = \alpha \times (t)$ in the Jukes-Cantor model. The diagonals of the rate matrix (no change events) are equal to the negative sum of the other entries which have a total rate of 3α for non-identical substitutions.

The branch lengths are considered as distances between pairs of sequences/nodes (extant taxa which are observed) and the ancestral nodes exists for each pair as a bifurcation. The ancestral node is denoted with i and the child nodes of i are j and k having diverged after a time t. If the rate of substitution from i to its 2 children are not identical the probabilities for the independent substitutions are $p_{i,j}(t)p_{i,k}(t)$, and the probability $p_{i,j}(t)^2$ when the rates are the same from the time of divergence from node i.

We consider the rate 3α because it is this total rate which affects the observed distances (ie. substitutions leaving an observable change). Node i is the common ancestor assumed to be in the middle of the length from node j to node k so the total time that has passed between the children nodes is $2t$. The average number of substitutions per position is $w_i = 3\alpha \times (2t) = 6\alpha t$. All the changes are equally likely, and the changes to equal nucleotides do affect the branch lengths unlike parsimonious construction methods which ignore homogeneous substitutions. Even though equal substitutions are not observed, they must be considered or else the branchlegnths will not be accurate. The probability for a nucleotide to be the same as the ancestral nucleotide is the no change probability, I used here to denote the probability there being no change in the entry. Each case of substitutions resulting in no change is considered to find I,

$$I = p_{j\to A}(t)p_{k\to A}(t) + p_{j\to C}(t)p_{k\to C}(t) + p_{j\to G}p_{k\to G}(t) + p_{j\to T}p_{k\to T}(t)$$

which is simplified assuming that the rate of mutation between the two species from node i is identical,

$$I = p_{j\to A}(t)^2 p_{j\to C}(t)^2 p_{j\to G}(t)^2 p_{j\to T}(t)^2. \tag{A.2}$$

The choice of the nucleotide to consider j is allowed for all 4 types and have those probabilities put into the equation. One of the terms will be the no change probability. Choosing an arbitrary initial nucleotide, eg. $j = A$, the probability for identical substitutions can be found

$$\begin{aligned} I &= \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)^2 + 3\left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)^2 \\ &= \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}. \end{aligned} \tag{A.3}$$

This last expression is useful for then finding the probability of non-identical substitutions, p, as $p = 1 - I$;

$$\begin{aligned} p &= 1 - I \\ &= 1 - \frac{1}{4} - \frac{3}{4}e^{-8\alpha t} \\ &= \frac{3}{4} - \frac{3}{4}e^{-8\alpha t} \\ \implies 8\alpha t &= -\ln(1 - \frac{4}{3}p). \end{aligned} \tag{A.4}$$

As said previously the rate of change is $6\alpha t$ and this result is $^4/_3$ larger, and therefore this result is scaled appropriately. The probability p is found by the fraction of changes to number of nucleotides. If n is the number of nucleotides and q the number of changes, $p = ^q/_n$. Now the expected distance can be calculated for the Jukes-Cantor model,

$$d_{JC} = -\frac{3}{4}\ln(1 - \frac{4}{3}p). \tag{A.5}$$

This is the derivation for the Jukes-Cantor model which is a special case as there are many simplifications in the model which do not take into account important biological considerations such as the transition-transversion ratio or the stationary distribution of the nucleotides. Other derivations are more involved and this one exists to demonstrate the interpretation of the branch lengths as distances in the expected number of substitutions (identical or not identical).

A.3 Hasegawa-Kishino-Yano (HKY) nucleotide substitution model

This mode is from 1985 and is found in Hasegawa *et al.* (1985). It is more complex than the previous two models of Kimura Kimura (1981) and Jukes and Cantor

(1969). It is more complex than the Jukes-Cantor model in that it allows different rates of substitution between transitions and transversions, and is more complex than the Kimura model in that the restriction on equal base frequencies for the stationary distribution is relaxed. An unequal base frequency of π is allowed.

$$\mathbf{Q} = \begin{pmatrix} -\beta(\pi_C+\pi_T)-\alpha\pi_G & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & -\beta(\pi_A+\pi_G)-\alpha\pi_T & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & -\beta(\pi_C+\pi_T)-\alpha\pi_A & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & -\beta(\pi_A+\pi_G)-\alpha\pi_C \end{pmatrix} \tag{A.6}$$

is the HKY85 rate matrix. This matrix can be simplified slightly by noticing that two variable can be created to group that stationary probabilities of transversions for a given nucleotide ($\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$),

$$\mathbf{R} = \begin{pmatrix} -\beta(\pi_Y)-\alpha\pi_G & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & -\beta(\pi_R)-\alpha\pi_T & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & -\beta(\pi_Y)-\alpha\pi_A & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & -\beta(\pi_R)-\alpha\pi_C \end{pmatrix}. \tag{A.7}$$

The equation for the probability of the events in the matrix in terms of time t are given (which are analogous to eq 1.45-eq 1.47 which use the rate matrix in a Taylor series expansion in eq 1.11):

$$\tilde{f}_j(t) = \pi_j\left(1-e^{-\beta t}\right) \tag{A.8}$$

$$\tilde{g}_j(t) = \pi_j + \pi_j\left(\frac{1}{\Pi_j}-1\right)e^{-\beta t} + \left(\frac{\Pi_j-\pi_j}{\Pi_j}-1\right)e^{-\beta t} \tag{A.9}$$

$$\tilde{d}_j(t) = \pi_j + \pi_j\left(\frac{1}{\Pi_j}-1\right)e^{-\beta t} - \left(\frac{\pi_j}{\Pi_j}-1\right)e^{-\beta t} \tag{A.10}$$

Where $\tilde{f}_j$ is for a transversion and $\tilde{g}_j$ a transition and $\tilde{d}_j$ a nucleotide remains unchanged. The subscript j in $\tilde{f}_j$, $\tilde{g}_j$, $\tilde{d}_j$ and π_j denotes the nucleotide which will substitute the original nucleotide. π_j denotes the value of the equilibrium probability for nucleotide j. In the equations $\Pi_j = \pi_A + \pi_G$ if the nucleotide is a purine (a nucleotide of C or T), and $\Pi = \pi_C + \pi_T$ if the nucleotide is a pyrimidine (a nucleotide of A or G). From inspection it can be seen that these equations satisfy the boundary conditions at $t = 0$ and $t = \infty$. As an example we take the starting original nucleotide to be A, for a transversion $\tilde{f}_j(t=0) = 0$ and $\tilde{f}_j(t=\infty) = \{\pi_C, \pi_T\}$, for a transition $\tilde{g}_j(t=0) = 0$ and $\tilde{g}_j(t=\infty) = \pi_G$, and for the replacement with the same nucleotide $\tilde{d}_j(t=0) = 1$ and $\tilde{d}_j(t=\infty) = \pi_A$.

On the stationary distribution of the rate matrix the constraint on π from eq 1.20 allows there to be 3 free parameters for this vector. There are also 2

free parameters for the value of the transition transversion ratio giving 5 free parameters minus 1 for a total of 4 free parameters in total in total by fixing the transversion rate. This is defined in eq 1.29. If the stationary distribution is chosen to be the uniform one, $(P(A) = P(C) = P(G) = P(T) = 0.25)$, then the HKY85 model reduces to the Kimura model. And this is the model used in most of the simulations to follow. The HKY model reduces to the Felsenstein model, Felsenstein (1981), when the transition-transversion parameters α and β are equal, but since this model is not used anywhere in this thesis no more is said about it.

A.4 Beta Distribution

The beta distribution is a continuous probability distribution with the domain on the interval $[0,1]$. There are two parameters α and β which govern the shape of the distribution in this interval. It is applied to random events such as Bernoulli trials where there are two possible outcomes. The α and β parameters represent one of the two possible events. For the modelling of coin tosses the α parameter can denote the number of times the coin fell on 'heads' and β for the number of times the coin was observed to fall on the 'tails' side. The Beta distribution differs from the binomial distribution where the value of the probability for a certain event is known. Here the distribution is over the proportion of the two possible events is evaluated. The distribution of the probability p of an event occurring makes it useful in Bayesian statistics for estimating the uncertainty. The posterior probability of p corresponding to $\alpha - 1$ positive events and $\beta - 1$ negative events with $1 - p$ probability is found via the Beta distribution. The pdf for the Beta distribution is,

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \tag{A.11}$$

where Γ is the gamma function

$$\Gamma(x) = \begin{cases} (x-1)! & x \in \mathbb{N} \\ \int_0^\infty t^{x-1}e^{-t}dt & x \in \mathbb{R}. \end{cases} \tag{A.12}$$

The gamma functions in the pdf of the beta distribution are abbreviated with $1/\mathrm{B}(\alpha,\beta)$,

$$f(x;\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}. \tag{A.13}$$

A.5 The effect on the transition probability due to the introduction of an extreme rate state

With the topology states the set of possible values is discrete. The rate states ρ can take on a continuous range of values within the valid domain of $[0,\infty)$ (negative rate state values are not defined). Not all of the rate state values will be applicable to the data and some can be extreme in that they create relatively much lower emission values for the data. Ratefactors are referred to in the ratefactor vector with a subscript index and the extreme valued ratefactor is referred to as ρ_∞ to identify it as a ratefactor with an extreme value rather that it taking position ∞ in the ratefactor vector ρ.

There is a set of ratefactors which can take on one of two states,$\rho_i \in \{\rho_1, \rho_{\text{inf}}\}$, where one element is an extreme rate state value. It is shown how the introduction of the extreme valued rate state affects the transition probability, $P(\mathbf{R}_t|\mathbf{R}_{t-1}$. For a rate state sequence, $P(\mathcal{D}|\nu_R) = \sum_{\mathbf{R}} P(D|\mathbf{R},\nu_R)P(\mathbf{R}|\nu)$ the addition of an extreme rate state will effect the transition probability and the probability of the sequence (significantly since such transitions are very unlikely). There are two cases to consider where the emission probability equals 1 and when it equals 0. These two cases exist since when the ratestate sequence contains only the relevant ratefactor states, $\mathbf{R} = (\mathbf{R}_1 = \rho_1, ..., \mathbf{R}_1 = \rho_1)$, where there are no other relevant ratefactors the probability of the data given the transition probability is 1. The 0 value exists in any other ratestate trajectory containing the extreme value rate state (all sequences containing a transition into the rate state with an extreme value). The sequence probability is,

$$P(\mathbf{R} = \rho_1, ..., \mathbf{R} = \rho_1|\nu) = \nu^{N-1}. \tag{A.14}$$

The same result is obtained from the emissions in the HMM states with the transitions where the rate state can take on either of two values in the set, $\mathbf{R}_t \in \{\rho_1, \rho_\infty\}$.

$$\prod_{t=1}^{N-1} \sum_{\mathbf{R}_t} P(y_t|\mathbf{R}_t)P(\mathbf{R}_t|\mathbf{R}_{t-1}) \tag{A.15}$$

which equals 0 in any case where $\mathbf{R}_t = \rho_{\text{inf}}$, and results in ν^{N-1} for the positions holding the proper ratefactor,

$$\prod_{t=1}^{N-1} (1)P(\mathbf{R}_t|\mathbf{R}_{t-1}) = \nu^{N-1} \tag{A.16}$$

In the case where there are now 3 rate states in the set to choose from $\rho_i \in \{\rho_1, \rho_2, \rho_{\text{inf}}\}$, with only one state being of an extreme value,

$$\sum_{\mathbf{R}} P(\mathcal{D}|\mathbf{R}, \nu) P(\mathbf{R}|\nu), \tag{A.17}$$

is used again where there are two cases; where the sequence equals 1 and when it equals 0. If $\forall_{i=1}^{N} \mathbf{R}_i \neq \rho_\infty$ it is 1 and 0 if $\exists \mathbf{R}_i = \rho_\infty$. For the case that it is 0 and there are transitions into the extreme state at some site i,

$$P(\mathbf{R}_{t \neq i} \neq \rho_\infty, ..., \mathbf{R}_{t=i} = \rho_\infty|\nu) = \nu^{N-k} \left(\frac{1-\nu}{2}\right)^{k-1}, \tag{A.18}$$

where k is the number of rates states having transitioned into a non-identical state. In the case where there are no transitions into the extreme value state, $\mathbf{R}_{i,t} \ni \rho_\infty$, there is this expression:

$$P(\mathbf{R}_1 \neq \rho_\infty, ..., \mathbf{R}_N \neq \rho_\infty|\nu) = \left(\frac{1-\nu}{2} + \nu\right)^{N-1} = \left(\frac{1+\nu}{2}\right)^{N-1}. \tag{A.19}$$

Here $(1-\nu)/2 + \nu$ is the total probability of not transitioning into the extreme valued state among the 3 possibilities. This shows that the amount of data increases (size of N), ν must also increase more towards the value of 1 to avoid a low compound probability since the alternative case results in extremely low emission values.

This working out shows that with the introduction of an extreme rate state, the transition probability will increase to penalise transitions into the extreme rate state value which would bring down the likelihood of the complete data sequence. As a result in the case where there are only 2 ratefactors with 1 being of a extreme value no consequence on the proper modelling of the data is made. In the case where there are 3 ratefactors with only 1 ratefactor being of an extreme value, the transition probability parameter increases with the amount of data hindering the model to make proper transitions between the 2 applicable ratefactor values of ρ_1 and ρ_2.

A.6 Optimising the hidden state variables with the Viterbi Algorithm

In *dynamic programming* the Viterbi algorithm is used for finding the most likely sequence of hidden states (*decoding*). A distribution is not returned, but a single

sequence is, which is the result of the optimisation process. The optimised result returned is called the *Viterbi* path. This algorithm is used in HMMs and can be used to find the most likely trajectory (path) of hidden topology states along the topology state HMM. The algorithm examines all possible paths towards the most likely one and keeps only the most likely. The result can be seen via a *trellis* diagram which is the trajectory from the first hidden state with transitions until the final hidden state.

The algorithm uses the eq 1.22or the relevant likelihood function, to use as a criteria in finding an optima. In this case the Viterbi algorithm searches for the topology state seq, $\mathbf{S}$, which maximises the joint probability of the DNA sequence alignment with $\hat{\mathbf{S}}$ as the optimal state sequence. The algorithm uses the emission probability model eq 1.72 and the transition state probabilities subfigure b) 1.18(costs) incrementally for finding the lowest penalising path through the state sequences. For the eq 1.87 the Viterbi algorithm predicts the recombinant regions and break points with,

$$\hat{\mathbf{S}} = \text{argmax}_{\mathbf{S}} P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \theta, \nu_S) \tag{A.20}$$

$$= \text{argmax}_{S,\ldots,S_N} P(S_1, \ldots, S_N|\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{w}, \theta, \nu_S). \tag{A.21}$$

The marginal probability of the first state $P(S_1)$ is found from the prior distribution of initial states, and the rest of the parameters ($\mathbf{w}$,θ, and ν_S) are estimated as is discussed later but assumed to be constant during the Viterbi algorithm and other state space optimising algorithms. From the dependency structure of eq 1.69 a recursion relationship is obtained allowing for the less than exponential running time of an exhaustive search. For the recursion operation the last state is considered, S_N, from this state the backtracking is done until the first state, S_1, which has no predecessors and there is a prior distribution over the initial states. The recursive relationship uses the log likelihood of the joint probability of the data and the intermediate variable $\gamma_t(S_t)$ to store the optimised search path up to the index t,

$$\gamma_N(S_N) = \max_{S_1,\ldots,S_{N-1}} \log P(\mathbf{y}_1, \ldots, \mathbf{y}_N, S, \ldots, S_N). \tag{A.22}$$

The site independence assumption made about the columns of DNA allow the log probability to write the product of the site probabilities in terms of sums (which

is defined in eq 1.73 and relaxed by the HMM),

$$\gamma_N(S_N) = \max_{S_1,\ldots,S_{N-1}} \left[\Sigma_{t=1}^{N} \log P(\mathbf{y}_t|S_t) + \Sigma_{t=2}^{N} \log P(S_t|S_{t-1}) + \log P(S_1))\right]. \tag{A.23}$$

From this previous expression the emission probability for the column of the index $t = N$ we are examining can be extracted from the sum of the logs,

$$\begin{aligned}\gamma_N(S_N) = \log P(\mathbf{y}_N|S_N) + \max_{S_{N-1}} \Big[\log P(S_N|S_{N-1}) + \\ \max_{S_1,\ldots,S_{n-2}} \left[\Sigma_{t=1}^{N-1} \log P(\mathbf{y}_t|S_t) + \right. \\ \left. \Sigma_{t=2}^{N-1} \log P(S_t|S_{t-1}) + \log P(S_1))\right]\Big].\end{aligned} \tag{A.24}$$

This expression separates the trellis until the present state being considered, so that an expression containing the previous states $1,\ldots,N-1$ can be made.

$$\gamma_N(S_N) = \log P(\mathbf{y}_N|S_N) + \max_{S_{N-1}} \left[\log P(S_N|S_{N-1}) + \gamma_{N-1}(S_{N-1})\right] \tag{A.25}$$

where the second term is used to obtain the recursive relationship. Since the state sequence which maximises the probability conditional on the data is equal to the same state sequence maximising the joint probability, the above recursion relationship can be used to find the posterior maximum of the state sequence given the data;

$$\begin{aligned}\max_{S_1,\ldots,S_N} P(S_1,\ldots,S_N|\mathbf{y}_1,\ldots,\mathbf{y}_N) &= \max_{S_1,\ldots,S_N} \\ &= \max_{S_N} \gamma_N(S_N).\end{aligned} \tag{A.26}$$

The backtracking from the recursive relationship finds $P(\hat{\mathbf{S}}|\mathcal{D}) = P(\hat{\mathbf{S}}_1,\ldots,\hat{\mathbf{S}}_N|\mathbf{y}_1,\ldots,\mathbf{y}_N)$. The algorithm starts with $\hat{S}_N = \operatorname{argmax}_{S_N} \gamma_N(S_N)$ and continues until the first state, with the relationship from eq A.25,

$$\hat{S_{N-1}} = \operatorname{argmax}_{S_{N-1}} \left[\log P(\hat{S}_N|S_{N-1}) + \gamma_{N-1}(S_{N-1})\right], \tag{A.27}$$

The exhaustive search of searching through every possible path would be exponential in terms of the number of states (columns in the sequence alignment). This recursive relationship (Viterbi algorithm), has a runtime of $O(K^3)$, which comes from the number of state values each state can take and the number of state transitions between those states which we assume to all be possible (in some cases not every state can transition to any other possible state). With the recursive relationship being incremental the algorithm is linear in N.

This algorithm demonstrates how the Markov assumption allows a more efficient method of inferring the state spaces than what would have arisen from a naive approach not utilising the structure of the model.

A.7 The Forward Algorithm

The Viterbi algorithm produces the most likely state sequence and the forward algorithm is similar, but there are differences. It computes the distribution of the hidden states given the set of observations from the beginning and ending at the same state index (*filtering*). With the states that are hidden represented by S_t, and t is the subscript taking values $1 \leq t \leq N$ as an index for the states. y_t are as before the observation variables which are not hidden, and $y_{1,\ldots,N}$ is the complete set of observations. The objective of the forward algorithm is to compute the α_t variables. These store the distribution over the hidden state variables S_t at site t for the observations from y_1 to y_t;

$$\alpha_t(S_t) = P(S_t, y_1, \ldots, y_t). \tag{A.28}$$

These are found recursively from the initial α variable computed for the first site $t = 1$, so for α_1 we consider the prior over the state variables π because there are no transitions from hidden states before it;

$$\alpha_{t=1} = \pi \times P(y_1|S_1). \tag{A.29}$$

For the set of observations from 1 to N, the set of observation variables are $y_1, \ldots, y_N$ and the forward algorithm can compute the probability of observing these observations;

$$P(y_1, \ldots, y_t) = \Sigma_i^K \alpha_t(S_i), \tag{A.30}$$

for the arbitrary variable index t and therefore the complete observation sequence can be found via

$$P(y_1, \ldots, y_N) = \Sigma_{i=1}^K \alpha_N(S_i). \tag{A.31}$$

The iteration $t+1$ from the previous states t is found by utilising the Bayesian network properties defined in subsection 1.9. This is where the α values storing the distribution of the hidden state values for the set of observation seen till that index are used. With the distribution over the state values the previous hidden state values and observations are independent given the alpha values. Eq 1.68 defines this independence of the states as is also depicted in the graphical model of Figure 1.15 (where the nodes represent the state variables and the arrows the parent/child relationships). With this the alpha values for the state variables in conjunction with the state transition value $P(S_{t+1}|S)$ and the observation probability $P(y_{t+1}|S_{t+1})$ (as defined in eq 1.69) the present alpha can be computed with

these 2 terms and the previous alpha value,

$$\alpha_t(S_i) = P(y_{t+1}|S_{t+1})\Sigma_{i=1}^{K} P(S_{t+1}|S_t = i)\alpha_{t-1}(S_t = i). \tag{A.32}$$

This way the complete distribution of alpha values at the site t can be found.

When starting from the complete set of observations the alpha parameters are found via recursion. The algorithm works backward utilising eq A.32, and the base case from which the recursion begins to return is eq A.29. Here is shown the working for the recursion relationship when beginning with the alpha at site N when none of the previous ones are already known:

$$\begin{aligned}
P(\mathbf{y}_1,\ldots,\mathbf{y}_N) &= \Sigma_{S_N}\alpha_N(S_N) && \text{(A.33)}\\
\alpha_n(S_n) &= P(\mathbf{y}_1,\ldots,\mathbf{y}_n,S_n) && \text{(A.34)}\\
&= \Sigma_{S_1}\ldots\Sigma_{S_{n-1}}P(\mathbf{y}_1,\ldots,\mathbf{y}_n,S_1,\ldots,S_{n-1},S_n) && \text{(A.35)}\\
&= \Sigma_{S_1}\ldots\Sigma_{S_{n-1}}P(S_t|S_{t-1},\nu_S)\prod_{t=1}^{n}P(\mathbf{y}_t|S_t) && \text{(A.36)}\\
&= P(\mathbf{y}_n|S_n)\Sigma_{S_{n-1}}P(S_n|S_{n-1})\Sigma_{S_1}\ldots\Sigma_{S_{n-2}}\prod_{t=1}^{n-1}P(\mathbf{y}_t|S_t)P(S_t|S_{t-1}) && \text{(A.37)}\\
&= P(\mathbf{y}_n|S_n)\Sigma_{S_{n-1}}P(S_n|S_{n-1})\alpha_{n-1}(S_{n-1}). && \text{(A.38)}
\end{aligned}$$

At the completion of the N steps of the algorithm (for each observation in the length of the HMM), the alpha value can be found, α_t giving the probability of state t being observed after the $y_1,\ldots,y_t$ observations were seen in the sequence.

The *forward algorithm* is an algorithm of significant importance. An alternative approach is the naive exhaustive search. To calculate the probability of the set of observations $P(\mathbf{y}_1,\ldots,\mathbf{y}_N)$, as in eq A.33, we would need to marginalise over the complete hidden statespace with nested sums. With K being the number of values the hidden state can take on (we also assume that the possible state values is homogeneous along the whole alignment), the nested sums create a computational complexity of K^N; $\Sigma_N\ldots\Sigma_1 K$. This is exponential in the running time due to the exponent being proportional to the length of the data set (observation length). This is not possible to use in realistic scenarios where observation lengths would result in excessive computational demand. From eq A.32 we can see that in total there will be N iterations (N recursions to be more accurate since we begin from eq A.33 at $t = N$), and there is a sum over the state space of the previous alpha and the present alpha needs to have this done for each hidden state value so there are 2 nested sums. This creates a run time proportional to K^2N which is

linear in terms of the amount of observational data. The state space complexity remains quadratic resulting in a much more optimised inference procedure than the exhaustive search.

A.8 Forward-backward Algorithm

The Forward algorithm, subsection A.7 calculated the probability of the observations from the first y_1 until the index t which could take on the value N, $t = N$, to compute the probability of the dataset with all the observations. The alphas computed could be used to examine the distribution of the hidden state values given the set observations before the index t. The alphas give the distribution of the hidden state values at that index value with all the previous observations and the previous hidden states marginalised over. It does *filtering*, and the forward-backward algorithm does *smoothing*. This algorithm singles out a single hidden state at a specified index t as does the forward algorithm utilising the observations following y_t as well which are the complete set until index $t = N$. This distribution we are interested in is,

$$P(S_t|y_1,\ldots,y_N). \tag{A.39}$$

Using Bayes rule we can write this distribution of the posterior of a hidden state given the data with the 3 terms, likelihood, prior, and marginal likelihood of the data;

$$P(S_t|y_1,\ldots,y_N) = \frac{P(y_1,\ldots,y_N|S_t)P(S_t)}{P(y_1,\ldots,y_N)}. \tag{A.40}$$

Given the HMM structure shown in figure 1.15 we can see that conditioning on a hidden state S_t can separate $P(y_1,\ldots,y_t|S_t)$,

$$= \frac{P(y_1,\ldots,y_t|S_t)P(y_{t+1},\ldots,y_N|S_t)P(S_t)}{P(y_1,\ldots,y_N)}.$$

The prior on the state $P(S_t)$ is used to cancel out one of the conditional distribution's denominator,

$$= \frac{P(y_1,\ldots,y_t,S_t)P(y_{t+1},\ldots,y_N|S_t)}{P(y_1,\ldots,y_N)}.$$

The term on the left is what is computed by the alpha values, eq A.32, and what is on the right-hand side of the numerator $P(y_{t+1},\ldots,y_N|S_t)$ is the beta value. The beta values used in the forward algorithm are computed from the recursive

relationship similarly to the defined alpha variables in eq A.32,

$$\beta_t = \Sigma_{i=1}^{K} \beta_{t+1}(S(i)) P(S(i)|S_t) P(y_{t+1}|S(i)), \tag{A.41}$$

that computes the distribution

$$P(y_{t+1}, \ldots, y_N|S_t). \tag{A.42}$$

The initialisation that this recursion step leads to is the base case of β_N which is,

$$\beta_N = 1. \tag{A.43}$$

Using the alpha and beta values at index t allow eq A.39 to be computed;

$$P(S_t|y_1, \ldots, y_N) = \frac{\alpha_t \beta_t}{P(y_1, \ldots, y_N)}. \tag{A.44}$$

This distribution is normalised over the complete hidden state space choices

$$1 = \Sigma_{i=1}^{K} P(S_t = i|y_1, \ldots, y_N) = \frac{\Sigma_{i=1}^{K} \alpha_t(S_i)\beta_t(S_i)}{P(y_1, \ldots, y_N)}. \tag{A.45}$$

And as a note this shows that the numerator on the right $\Sigma_{i=1}^{K} \alpha_t(S_i)\beta_t(S_i)$ is equal to the probability of the complete set of observations.

We define now gamma variables γ_t over the hidden state variables as being the marginal posterior distribution of the hidden states given the complete set of observation data,

$$\gamma_t = P(S_t|y_1, \ldots, y_N) \tag{A.46}$$

which is found from product of the alpha and beta values at that site:

$$\gamma_t(S_i) = \alpha_t(S_i)\beta(S_i) \tag{A.47}$$

This is going to be very important when examining the marginal posterior distribution across the hidden states for the topologies and ratefactors at each site in the alignment. Plots will be produced from simulations showing the marginal posterior distribution along the sites for the hidden state values.

A.9 Nested Gibbs-within-Gibbs for the HMM hidden state space sampling

Here is described an alternative method to the forward-backward algorithm described in subsection A.8 for finding the hidden state trajectory along the HMM

by sampling state sequences. The method of nested Gibbs-within-Gibbs infers the hidden state sequence of the HMM by sampling from the model in a Bayesian way with the samples being taken from the posterior distribution. This approach of the nested Gibbs-within-Gibbs scheme is described and implemented in Husmeier and McGuire (2003) for the purpose of sampling the state sequence $\mathbf{S}$. The method is discussed previously in Robert *et al.* (1993) where the authors suggest this approach of sampling each state S_t individually conditioned on the other states. Each Gibbs step draws a state conditional on the others in the following scheme (where i is the iteration number of the Gibbs simulation):

$$\begin{aligned}
S_1^{(i+1)} &\sim P(\cdot|S_2^{(i)},\ldots,S_N^{(i)},\mathbf{R}^{(i)},\nu_S^{(i)},\nu_R^{(1)},\mathbf{w}^{(i)},\rho^{(i)},\mathcal{D}) \\
S_2^{(i+1)} &\sim P(\cdot|S_1^{(i+1)},S_3^{(i)},\ldots,S_N^{(i)},\mathbf{R}^{(i)},\nu_S^{(i)},\nu_R^{(1)},\mathbf{w}^{(i)},\rho^{(i)},\mathcal{D}) \\
&\vdots \\
S_N^{(i+1)} &\sim P(\cdot|S_1^{(i+1)},S_2^{(i+1)},\ldots,S_{N-1}^{(i+1)},\mathbf{R}^{(i)},\nu_S^{(i)},\nu_R^{(1)},\mathbf{w}^{(i)},\rho^{(i)},\mathcal{D}). \quad \text{(A.48)}
\end{aligned}$$

The computational complexity of sampling each state conditionally on the rest of the states is greatly reduced by the sparseness of the connectivity in the HMM structure. The HMM structure for this sampling scheme reduces the sampling to be conditional on the nodes to the left and right of the hidden state of concern. This is due to the Markov blanket, comprising of the surrounding nodes being the parents, children, and coparents as explained in Heckerman (1999). For sampling on a particular hidden topology state S_t at site t,

$$P(S_t|S_1,\ldots,S_{t-1},S_{t+1},\ldots,S_N,\mathbf{R},\nu_S,\nu_R,\mathbf{w},\rho,\mathcal{D}), \quad \text{(A.49)}$$

this is simplified to (via the Markov Blanket),

$$\begin{aligned}
&= P(S_t|S_{t-1},S_{t+1},\mathbf{R},\nu_S,\nu_R,\mathbf{w},\rho,\mathbf{y}_t) \quad \text{(A.50)} \\
&\propto P(S_{t+1}|S_t,\nu_s)P(S_t|S_{t-1},\nu_s)P(\mathbf{y}_t|S_t,\mathbf{R},\mathbf{w},\rho,\nu_R) \quad \text{(A.51)}
\end{aligned}$$

and this last expression can be normalised to give a probability which can be sampled from easily since the set of topologies is a finite discrete set.

A.10 Importance Sampling

Importance sampling generates independent samples from a target distribution using a different distribution to propose values to sample points from. It is used

in cases where the target distribution cannot be used to generate samples from directly. The ability to generate samples from the distribution can be used to calculate the volume of a distribution or the expected value of a distribution over a particular set of dimensions for example.

The samples are iid taken from a distribution different from the target distribution for importance sampling as the target distribution cannot have samples directly taken. The proposal distribution $g(x)$ is used to generate points from and can differ from a uniform distribution and the volume does not change the final outcome but only relative 'importance' given to points based on the density of the distribution. The distribution of concern is $f(x)$, and the integral can be found by,

$$\int_a^b \frac{f(x)}{g(x)} g(x) dx.$$

$w = \frac{f(x)}{g(x)}$ represents the weighting of the target distribution to the distribution samples are proposed from scaling the number of times this point should have been proposed from. This continuous distribution is approximated by a discrete sum of the samples which is average over and to obtain the expected value the total scaling values must be compensated for:

$$\mathrm{E} = \frac{\sum_{i=1}^{N} f(x_i) w(x_i)}{\sum_{j=1}^{N} w(x_j)}. \tag{A.52}$$

Where N is the number of samples taken.

The choice of $g(x)$ must be reasonable in that it cannot have a zero density value where $f(x) > 0$. Large differences in the values will cause poor convergences, eg. where $g(x_i)$ has a relatively low value samples from $f(x_i)$ regardless if they are large will be rarely sampled and the regions of large mass may be ignored.

Importance sampling is used as part of annealed importance sampling discussed in chapter 3.

A.11 Details of the Gibbs sampling scheme used for the improved phylogenetic FHMM

We briefly describe the Gibbs sampling procedure that we used for the improved phylogenetic FHMM described in Section 2.12.3.

We sampled the hidden state sequences and model parameters according to the Gibbs sampling scheme described in Sections 2.7 and 2.12. We carried out 200

Gibbs sampling steps in the burn-in phase, and 200 steps in the sampling phase. Recall that each Gibbs sampling step includes a set of Metropolis-Hastings (MH) steps for adapting the branch lengths, according to equations (2.48) and (2.51). Within each Gibbs step, we carried out 200 MH steps for the MH burn-in phase, and 1200 MH steps for the MH sampling phase. The final branch length vector was kept, and constituted the output of the Gibbs sampling step (2.48). During the MH burn-in phase, the parameter α of the proposal distribution (3.2.1) was adjusted, as described in Section 2.12.3. We used the MH sampling phase to compute, for all branch lengths, the potential scale reduction factor of Gelman and Rubin (1992).

For the simulations thus carried out, we found that the potential scale reduction factor was consistently smaller than 1.1, indicating a satisfactory degree of convergence. The marginal posterior probabilities of the topology states, $P(S_t = \Psi_k|\mathcal{D})$, were computed straight from the state sequences $\{\mathbf{S}_i\}$ sampled during the sampling phase of the Gibbs sampling scheme by application of (2.44); the results are shown in Figures 2.8 and 2.9.

A.12 Transformation of Random variables

The transformation of random variables for continuous probability distributions densities is than in the case of discrete distributions. For the univariate case of a transformation from a density function $P(x)$ to another density function $P(y)$ the equality is based on this equation,

$$P(x) = P(y)\frac{dy}{dx}. \tag{A.53}$$

The concept of conserving probability during a transformation is depicted in Figure A.1. In the multivariate case where the density functions cover more than one dimension ($f(x,y)$ to $g(r,\phi)$), the Jacobian is used. We switch from

$$f(x,y) = g(r,\phi)\,|\det \mathbf{J}|\,, \tag{A.54}$$

and the modulus of the Jacobian matrix determinant is used to avoid negative values. To provide a more in depth explanation an example will be used. The transformation of the function $g(r,\phi)$ to the function $f(x,y)$ where r and ϕ are continuous independently distributed random variables. The variable ϕ has a

uniform distribution over the interval $[0,2\pi]$, and the variable r is uniformly distributed over the interval $[0,10]$. The variables x and y are defined as $x = r\cos(\phi)$ and $y = r\sin(\phi)$. The probability density for the distribution of x and y is given by the transformation using the Jacobian. This is represented with the partial derivatives as,

$$f(x,y) = g(r,\phi)\left|\frac{\partial(r,\phi)}{\partial(x,y)}\right|. \tag{A.55}$$

For convenience the inverse of the Jacobian can be used and the reciprocal of the value is taken,

$$f(x,y) = g(r,\phi)\left|\frac{\partial(x,y)}{\partial(r,\phi)}\right|^{-1}. \tag{A.56}$$

The expansion of the partial derivatives is,

$$f(x,y) = g(r,\phi)\begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \phi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \phi} \end{vmatrix}^{-1}. \tag{A.57}$$

Substituting the equations and performing the differentiations,

$$f(x,y) = g(r,\phi)\begin{vmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{vmatrix}^{-1} = g(r,\phi)\frac{1}{r\left(\cos^2\phi+\sin^2\phi\right)} = g(r,\phi)\frac{1}{r}. \tag{A.58}$$

The functions are set to equal each other with $f(x,y) = g(r,\phi)\times r^{-1}$. This value of r can be expressed in terms of x and y as $r = \sqrt{x^2+y^2}$, and for ϕ it is given by $\phi = \arctan\frac{y}{x}$. The distribution for $g(r,\phi)$ is the product of the independent uniform distributions over their intervals,

$$g(r,\phi) = \frac{1}{10}\times\frac{1}{2\pi}\times\mathbb{I}(0\le r\le 10)\times\mathbb{I}(0\le\phi\le 2\pi), \tag{A.59}$$

where $\mathbb{I}$ is the indicator function which is 1 when the arguments are true and 0 otherwise. This gives the formula for $f(x,y)$ as,

$$f(x,y) = \frac{g(\sqrt{x^2+y^2},\arctan(\frac{y}{x}))}{r} = \frac{1}{10}\times\frac{1}{2\pi}\times\frac{1}{r} = \frac{1}{20\pi r}. \tag{A.60}$$

To demonstrate the use we chose the values of $x = 2$ and $y = 0$ and find the probability of $P(x,y)$ which is,

$$P(x=2,y=0) = \frac{1}{20\pi\times\sqrt{2^2+0^2}} = \frac{1}{40\pi}. \tag{A.61}$$

To do the same, with the previous example, over the discrete domain there is no use of the Jacobian as each value in the domain is a probability rather than a

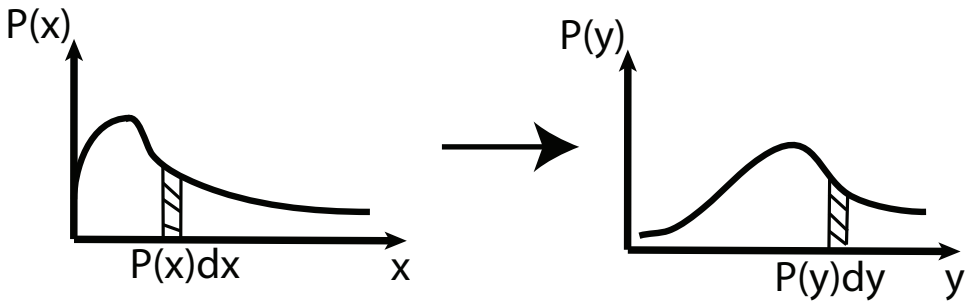


Figure A.1: Transformation of random variables

During transformations of random variables the probability has to be conserved. Probability for an infinitesimal area $P(x) = P(y)\frac{dy}{dx}$ must be equal in the transformation, which is shown in the shaded area in the graph above. The infinitesimal probability is the infinitesimal area under the probability density function $P(x)dx$ and the conservation of probability implies that $P(y)dy = P(x)dx$, leading to $P(y) = P(x)\left[\frac{dx}{dy}\right]$. In the multivariate case $\left[\frac{dx}{dy}\right]$ becomes the modulus of the determinant of the Jacobian.

density. Assume the distributions for r and ϕ are uniform over the same interval, equally separated, with a cardinality of 10 each. There are then 100 possible pairings which are uniformly distributed as well for $P(x,y)$,

$$P(x,y) = \frac{1}{100} \times \mathbb{I}(0 \leq r \leq 10) \times \mathbb{I}(0 \leq \phi \leq 2\pi). \tag{A.62}$$

This is over the valid set of $x = r\cos\phi$ and $y = r\sin\phi$ and 0 otherwise. For the variables taking on the same values as the example previously, $x = 2$ and $y = 0$ (these values appear in the support of the distribution), the result is $P(x = 2, y = 0) = \frac{1}{100}$.

A.13 Algorithm for the rate factor RJMCMC scheme

In the table 1, is given the pseudocode for RJMCMC scheme for the ratefactors.

A.14 Algorithm for the MCMC of the branch lengths and Codon relative rate vector

In the table 2, is given the pseudocode for the sampling of the branch lengths and vector of relative codon rate heterogeneity.

A.15 Ratefactor MCMC algorithm

In the table 3, is given the pseudocode for the Metropolis-Hastings sampling of the ratefactors.

A.16 The Jacobian for RJMCMC

Consider the simple case of a birth-death move, in which a component is added to the existing ones. Assume, for simplicity of exposition, that we only have one component, with parameter $\rho \in I$, where I is some interval of real numbers. In the birth move, we add a new component, $u \in I$, sampled from some distribution

Algorithm 1 RJMCMC for the rate states

function = $FHMM-RJMCMC(\mathbf{S}_{init}, \mathbf{R}_{init}, \mathbf{R}_{init}, \mathbf{w}_{init}, \theta_\theta, \lambda_{init})$

for $i = 1$ to Gibbs iteration end **do**

[**S**,**R**,loglik] ← FHMM(**S**,**R**,ρ,**w**,θ,λ) {Sampling the hidden state values of the topology and rates factors for the sites via the FHMM.}

[**w**,λ] ← **w**andλMCMC(**S**,**R**,ρ,**w**,λ)

b_k = c×minA$\{1, \frac{q(k+1)}{q(k)}\}$

d_k = c×minA$\{1, \frac{q(k-1)}{q(k)}\}$

m_k = 1 - b_k - d_k

r1 = rand()

if r1 $\leq b_k$ **then**

log[ρ′] ← sampleUniform$[min(-3), max(2)]$

ρ(1,k+1)' = ρ′

$\theta_{\nu_\mathbf{R}}$' = Reconfig$\nu_\mathbf{R}$[$\theta_{\nu_\mathbf{R}}$

[**R**',loglik'] ← FHMM(**S**,**R**,ρ',**w**,θ',λ)

accept = minA$\{1, \frac{loglik'}{loglik}\}$

r2=rand()

if accept ≥ r2 **then**

ρ= ρ'

R= ρ'

θ = θ'

end if

else if $b_k \leq r1 \leq d_k$ **then**

ρ' (delete random selected state)

$\theta_{\nu_\mathbf{R}}$' = Reconfig$\nu_\mathbf{R}$[$\theta_{\nu_\mathbf{R}}$

[**R**',loglik'] ← FHMM(**S**,**R**,ρ',**w**,θ',λ)

accept = minA$\{1, \frac{loglik'}{loglik}\}$

r2=rand()

if accept ≥ r2 **then**

ρ= ρ'

θ = θ'

R= ρ'

end if

else

[ρ] ← ρMCMC(**S**,**R**,ρ,**w**,λ)

end if

save the number of rate states and the values

θ' ← sample θ

end for

return $(\mathbf{S}, \mathbf{R}, \rho, \mathbf{w}, \theta, \lambda)$

Algorithm 2 MCMC for the nucleotide branch lengths and codon vector parameters $\lambda, \mathbf{w}$

function($\mathcal{D}, \mathbf{w}_{init}, L_{\mathbf{w}}, \lambda_{init}, L_{\lambda}, burninN, sampleN, \mathbf{S}, \rho, \mathbf{R}$)
$\alpha_{\mathbf{w}} = 3$
$\alpha_{\lambda} = 1$
for topo in TopologylNumber **do**
lik = SequenceProbability($topo, \mathbf{w}, \mathcal{D}, \mathbf{S}, \rho, \lambda, \mathbf{R}$)
$prior_{\mathbf{w}}$ = exponentialDist($\mathbf{w}, \alpha_{\mathbf{w}}$)
$prior_{\lambda}$ = exponentialDist($\lambda, \alpha_{\lambda}$) {the likelihood of the DNA sequence alignment given all theparameters for the nucleotide substitution matrix and the rate states for the parameters}
i=0
while i < burninN+sampleN **do**
i++
$[\mathbf{w}', P_{\mathbf{w}\leftarrow\mathbf{w}'}, P_{\mathbf{w}'\leftarrow\mathbf{w}}]$ = DirichletSample($\mathbf{w}, L_{\mathbf{w}}$)
$[\mathbf{w}', P_{\lambda\leftarrow\lambda'}, P_{\lambda'\leftarrow\lambda}]$ = DirichletSample(λ, L_{λ}) { priors for the new and old sets of the branch lengths and codon rate factor vectors}
$prior'_{\mathbf{w}}$ = exponentialDist($\mathbf{w}', \alpha'_{\mathbf{w}}$)
$prior'_{\lambda}$ = exponentialDist($\lambda', \alpha'_{\lambda}$)
lik' = SequenceProbability(topo,$\mathbf{w}', \mathcal{D}, \mathbf{S}, \rho, \lambda', \mathbf{R}$)
Accept = $\min\{1, \frac{prior'_{\lambda} \times prior'_{\mathbf{w}} \times lik' \times P_{\mathbf{w}\leftarrow\mathbf{w}'} \times P_{\lambda\leftarrow\lambda'}}{prior_{\lambda} \times prior_{\mathbf{w}} \times lik \times P_{\mathbf{w}'\leftarrow\mathbf{w}} \times P_{\lambda'\leftarrow\lambda}}\}$
r2 = rand()
if Accept $\geq$ r2 **then**
$\lambda = \lambda'$
$\mathbf{w} = \mathbf{w}'$
lik = lik'
end if
if i $\leq$ burninN **then**
if acceptance ≤ 0.30 **then**
$L_{\lambda} = \frac{L_{\lambda}}{2}$
$L_{\mathbf{w}} = \frac{L_{\mathbf{w}}}{2}$
else if acceptance ≥ 0.70 **then**
$L_{\lambda} = L_{\lambda} \times 2$
$L_{\mathbf{w}} = L_{\mathbf{w}} \times 2$
end if
end if
end while
end for
return $(\mathbf{w}, L_{\mathbf{w}}, \lambda, L_{\lambda})$

Algorithm 3 MCMC for the ρ

function($\mathcal{D}, \mathbf{w}, \lambda, burninN, sampleN, \mathbf{S}, \rho, \mathbf{R}$)
for topo in Topologies **do**
  for ρ_i in ρ **do**
    likρ_i = SequenceProbabilityR(topo,$\mathbf{w}, \mathcal{D}, \mathbf{S}, \rho_i, \lambda, \mathbf{R}$) {this function calculates the likelihood of the subsequences for which the ratestates have allocated this particular rate value}
  end for
  i=0
  while i < burninN+sampleN **do**
    i++
    for ρ_i in ρ **do**
      $\rho_i' = \rho_i \pm (rand[0,1] - 0.5) \times L_{\rho,i}$ {with reflection}
      likρ_i' = SequenceProbabilityR(topo,$\mathbf{w}, \mathcal{D}, \mathbf{S}, \rho_i', \lambda, \mathbf{R}$)
      Accept = $\min\{1, \frac{lik\rho_i'}{lik\rho_i}\}$
      r2 = rand()
      if Accept $\geq$ r2 **then**
        $L_{\rho,i} = L'_{\rho,i}$
      end if
      if i $\leq$ burninN **then**
        if acceptance ≤ 0.30 **then**
          $L_{\rho,i} = \frac{L_{\rho,i}}{2}$
        else if acceptance ≥ 0.70 **then**
          $L_{\rho,i} = L_{\rho,i} \times 2$
        end if
      end if
    end for
  end while
end for
return $(\rho_{1,N}, L_{\rho,i})$

over I. In the death move, we discard one of the components. This gives us the bijection

$$(\rho, u) = (\rho_1, \rho_2) \tag{A.63}$$

and the Jacobian is:

$$\begin{vmatrix} \frac{\partial \rho}{\partial \rho_1} & \frac{\partial \rho}{\partial \rho_2} \\ \frac{\partial u}{\partial \rho_1} & \frac{\partial u}{\partial \rho_2} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1. \tag{A.64}$$

As an alternative to the birth/death move, consider a merge/split move. Given the vector (ρ_1, ρ_2), rather than just discarding a component, we merge them: $\rho = (\rho_1 + \rho_2)/2$. In order to get a bijection, we introduce the following auxiliary variable: $u = (\rho_1 - \rho_2)/2$ This leads to the following bijection:

$$(\rho, u) = \left(\frac{\rho_1 + \rho_2}{2}, \frac{\rho_1 - \rho_2}{2} \right) \tag{A.65}$$

Expressing (ρ_1, ρ_2) as a function of ρ and u gives:

$$(\rho_1, \rho_2) = (\rho + u, \rho - u) \tag{A.66}$$

This leads to the following Jacobian:

$$\begin{vmatrix} \frac{\partial \rho_1}{\partial \rho} & \frac{\partial \rho_1}{\partial u} \\ \frac{\partial \rho_2}{\partial \rho} & \frac{\partial \rho_2}{\partial u} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} = |-2| = 2. \tag{A.67}$$

The conclusion is that simple birth/death moves have a Jacobian that is the unit matrix; hence, the determinant is 1 and does not make any contribution to the RJMCMC acceptance probabilities. For a merge/split move, the Jacobian us usually different from the unit matrix, and its determinant cannot be neglected in the RJMCMC acceptance ratio. While, at the face of it, merge/split moves might appear as more sophisticated than birth/death moves, a study by Boys and Henderson (2002) has shown that they do not lead to any improved mixing/convergence than the simpler birth/death moves. The latter were therefore chosen in the work of my PhD thesis.

Bibliography

Boys, R. J. and Henderson, D. A. (2002) A comparison of reversible jump MCMC algorithms for DNA sequence segmentation using hidden Markov models. *Computer Science and Statistics*, **33**, 35–49.

Boys, R. J., Henderson, D. A. and Wilkinson, D. J. (2000) Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Applied Statistics*, **49**, 269–285.

Casella, G. and George, E. I. (1992) Explaining the Gibbs sampler. *The American Statistician*, **46**, 167–174.

Celeux, G., Hurn, M. and Robert, C. P. (2000) Computational and inferential difficulties with mixture posterior distributions. *Journal of American Statistical Association*, **95**, 957–970.

Dirk Husmeier, S. R. E., Richard Dybowski (2003) *Probabilistic Modeling in Bioinformatics and Medical Informatics*. Springer.

Durbin, R., Eddy, S. R., Krogh, A. and Mitchison, G. (1998) *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.

Edwards, A. (1970) Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society B*, **32:2**, 155–174.

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

Felsenstein, J. (1978b) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, **27**, 401–440.

Felsenstein, J. (1978b) The number of evolutionary trees. *Systematic Zoology*, **27**, 27–33.

Felsenstein, J. (1981) Evolution trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.

Felsenstein, J. (1996) Phylip. Free package of programs for inferring phylogenies, available from http://evolution.genetics.washington.edu/phylip.html.

Felsenstein, J. and Churchill, G. A. (1996) A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, **13**, 93–104.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **Vol. 7, No.4**, 457–472.

Goldman, G. (1990) Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. *Systematic Zoology*, **39**, 345–361.

Grassly, N. C. and Holmes, E. C. (1997) A likelihood method for the detection of selection and recombination using nucleotide sequences. *Molecular Biology and Evolution*, **14**, 239–247.

Green, P. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

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

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Heckerman, D. (1999) A tutorial on learning with Bayesian networks. In Jordan, M. I. (ed.), *Learning in Graphical Models*, Adaptive Computation and Machine Learning, pp. 301–354. MIT Press, Cambridge, Massachusetts.

Hein, J. (1993) A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, **36**, 396–405.

Huelsenbeck, J., Ane, C., Larget, B. and Ronquist, F. (2008) A Bayesian perspective on a non-parsimonious parsimony model. *Systematic Biology*, **57:3**, 406–419.

Husmeier, D. (2005) Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models. *Bioinformatics*, **21**, ii166–ii172.

Husmeier, D., Dybowski, R. and Roberts, S. (2005a) *Probabilistic Modeling in Bioinformatics and Medical Informatics*. Advanced Information and Knowledge Processing. Springer, New York.

Husmeier, D. and Mantzaris, A. V. (2008) Addressing the shortcomings of three recent bayesian methods for detecting interspecific recombination in dna sequence alignments. *Statistical Applications to Genetics and Molecular Biology (SAGMB)*, **7**, 166–172.

Husmeier, D. and McGuire, G. (2002) Detecting recombination with MCMC. *Bioinformatics*, **18**, S345–S353.

Husmeier, D. and McGuire, G. (2003) Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Molecular Biology and Evolution*, **20**, 315–337.

Husmeier, D. and Wright, F. (2001) Detection of recombination in DNA multiple alignments with hidden Markov models. *Journal of Computational Biology*, **8**, 401–427.

Husmeier, D., Wright, F. and Milne, I. (2005b) Detecting interspecific recombination with a pruned probabilistic divergence measure. *Bioinformatics*, **21**, 1797–1806.

Huson, D. H. and Bryant, D. (2006) Application of Phylogenetic networks in Evolutionary Studies. *Molecular Biology and Evolution*, **23**, 254–267.

Jasra, A., Holmes, C. and Stephens, D. (2005) Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, **20**, 50–67.

Jukes, T. and Cantor, C. (1969) *Evolution of protein molecules*. Academic Press, New York.

Kimura, M. (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **Vol. 78**, pp. 454–458.

Kingman, J. F. C. (1982) The coalescent. *Stochastic Process Applications*, **13**, 235–248.

Larget, B. and Simon, D. L. (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, **16**, 750–759.

Lehrach, W. and Husmeier, D. (2009) Segmenting bacterial and viral DNA sequence alignments with a transdimensional phylogenetic factorial hidden Markov model. *Applied Statistics*, **in press**.

Lehrach, W. P. (2008) *Bayesian machine learning methods for predicting protein-peptide interactions and detecting mosaic structures in DNA sequence alignments*. Ph.D. thesis, Univeristy of Edinburgh.

Mantzaris, A. V. and Husmeier, D. (2009) Distinguishing regional from within-codon rate heterogeneity in dna sequence alignments. In *Pattern Recognition in Bioinformatics*.

Marin, J. M. and Robert, C. P. (2007) *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer New York.

Maynard Smith, J. (1992) Analyzing the mosaic structure of genes. *Journal of Molecular Evolution*, **34**, 126–129.

McGuire, G. and Wright, F. (2000) TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics*, **16**, 130–134.

McGuire, G., Wright, F. and Prentice, M. (2000) A Bayesian method for detecting past recombination events in DNA multiple alignments. *Journal of Computational Biology*, **7**, 159–170.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Minin, V. N., Dorman, K. S., Fang, F. and Suchard, M. A. (2005) Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, **21**, 3034–3042.

Neal, R. M. (1993) Probabilistic inference using markov chain monte carlo methods. Technical report, Dept. of Computer Science, University of Toronto.

Neal, R. M. (2001) Annealed importance sampling. *Statistics and Computing*, **11**, 125–139.

Papoulis, A. (1991) *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Singapore, 3rd edition.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA, USA.

Pickett, K. and Randle, C. (2005) Strange Bayes indeed: uniform topological priors imply non-uniform clade priors. *Molecular Phylogenetics and Evolution*, **34:1**, 203–211.

Raftery, A. E. (1996) Hypothesis testing and model selection. In Gilks, W. R., Richardson, S. and Spieglehalter, D. J. (eds.), *Markov Chain Monte Carlo in Practice*, pp. 163–187. Chapman & Hall, Suffolk. ISBN 0-412-05551-1.

Rambaut, N. C., A.; & Grassly (1996) Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*

Robert, C. P., Celeux, G. and Diebolt, J. (1993) Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statistics & Probability Letters*, **16**, 77–83.

Robertson, D. L., Sharp, P. M., McCutchan, F. E. and Hahn, B. H. (1995) Recombination in HIV-1. *Nature*, **374**, 124–126.

Suchard, M. A., Weiss, R. E., Dorman, K. S. and Sinsheimer, J. S. (2003) Inferring spatial phylogenetic variation along nucleotide sequences: A multiple

changepoint model. *Journal of the American Statistical Association*, **98**, 427–437.

Thompson, A. (1975) *Human evolutionary trees*. Cambridge University Press.

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

Velasco, J. (2008) The prior probabilities of phylogenetic trees. *Biology and Philosophy*, **23:4**, 455–473.

Webb, A., Hancock, J. and Holmes, C. (2009) Phylogenetic inference under recombination using Bayesian stochastic topology selection. *Bioinformatics*, **in press**.

Werhli, A. V., Grzegorczyk, M., Chiang, M. T. and Husmeier, D. (2006) *Improved Gibbs sampling for detecting mosaic structures in DNA sequence alignments*, pp. 23–24. Cnetro Internacional de Matematica, Coimbra, Portugal.

Wiuf, C., Christensen, T. and Hein, J. (2001) A simulation study of the reliability of recombination detection methods. *Molecular Biology and Evolution*, **18**, 1929–1939.

Yang, Z. (1995) A space-time process model for the evolution of dna sequences. *Genetics*, **139**, 993–1005.

Yang, Z. and Rannala, B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Molecular Biology and Evolution*, **14**, 717–724.

Zhou, J. and Spratt, B. G. (1992) Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. *Molecular Microbiology*, **6**, 2135–2146.