UNIVERSITAT POLITÈCNICA DE CATALUNYA
Facultat de Matemàtiques i Estadística.

Algebraic tools in phylogenomics.

Anna Magdalena Kędzierska

supervisors: Marta Casanellas i Rius and Roderic Guigó i Serra

Barcelona, Febrer de 2012

Contents

“Some men have thousands of reasons why they cannot do what they want to, when all they need is one reason why they can.” Marta Graham

Dla Dziadka Stefana i Aleksa

Acknowledgments

Interdisciplinary work is not a lonely effort. I would like to thank all the amazing people that helped and accompanied me through this process. These last years have been full of challenges, travels and immensely enriching experiences.

Kochani rodzice, dziękuję, że zawsze wspieracie mnie we wszystkich moich wyborach. Kinka, kochana sista, jesteś dla mnie ostoją i inspiracją. Dziękuję moim przyjaciołom za ich niesamowitą obecność: Jusi, Markowi, Aniuli, Sarze, Krzysztofowi i Sylwanie. Konie kraść i przenosić góry to z Wami bułka z masłem.

I would like to express deep gratitude to my exceptional supervisors, prof. Marta Casanellas and prof. Roderic Guigó, who created the possibility for this project and provided motivation and support for my growth. I am truly inspired by their openmindedness. I would also like to thank prof. Lior Pachter, prof. Bernd Sturmfels, prof. Mathias Drton and prof. Sonja Petrovic for inviting and hosting me in Berkeley and Chicago, and so openly sharing their expertise and knowledge.

I thank my colleagues at the PRBB: Sara, Colin, Julien, Marco, Andrea, Mireya, Andre, Hagen, Leszek, Darek, Pedro, Romina, Oscar and to all those who made my stay and work in Barcelona so pleasurable. Dono les gràcies als meus amics de la UPC, Abdó i Xavi, per totes les llargues discussions i per deixar-me ser el DJ de l'oficina. También querría agradecer a Nira, Salome, Deborah, Rodrigo y Joan, mis queridos amigos, por su soporte en los tiempos divertidos y los que no lo eran tanto. I thank Ishka, Stephane and prof. Martin Kuiper for their friendship, broadening my horizons and bringing the best out of me. Last but not least, I would like to thank my mentors: prof. Ryszard Komorowski, prof. John Hinde and dr Dirk Husmeier for encouraging me to feed my curiosity about the world through science.

I would like to credit the following grants: MTM2009-14163-C02-02 and Generalitat de Catalunya 2009 SGR 1284 and the FPI grant BES-2007-16623 of Ministerio de Ciencia e Innovación.

Introduction

Mathematics Is Biology's Next Microscope, Only Better; Biology Is Mathematics' Next Physics, Only Better

(Cohen, 2004)

The aim of this thesis was to provide a better understanding of evolutionary processes using tools from algebraic statistics. This thesis is therefore an interdisciplinary work that merges the areas of algebraic geometry, group theory, statistics, phylogenetics and genomics.

In phylogenetics, the goal is to reconstruct the ancestral relationships among organisms. Most of the widely used phylogenetic reconstruction methods are based on the mathematical models describing the molecular evolution of DNA along a *phylogenetic tree* $\mathcal{T}$. The leaves of the tree $\mathcal{T}$ are labeled by a set of currently living organisms and the interior nodes represent their common ancestors. Different shapes of the tree, called *tree topologies*, correspond to distinct speciation processes.

We will assume (as it is commonly done) that the sites in a DNA sequence are independent and identically distributed (*iid* hypothesis) and thus we model evolution one site at a time. An align collection (based on similarity) of the DNA sequences at the contemporary taxa is called a *multiple sequence alignment.* We will view it as an array, where the DNA sequences are placed in rows and the columns represent represent the evolution of a single character on $\mathcal{T}$.

The length of an edge in a phylogenetic tree is called a *branch length* and quantifies the amount of divergence between the species at its vertices. Branch lengths are measured in the expected number of substitutions per site that have occurred during the evolutionary process along that edge. A common way of modeling evolution is to consider a Markov process along $\mathcal{T}$. Namely, the states in the process correspond to the four different nucleotides, and the substitution matrices (or transition matrices) on each edge of the tree contain the probabilities of changes between nucleotides as the evolution proceeded along that edge (its entries correspond to the conditional probabilities, $P(x|y, e)$, that a nucleotide y at the parent node of e is substituted by nucleotide x at the child node). The parameters of a Markov evolutionary model are therefore the substitution matrices A^e assigned to each edge of the tree and, if a distinct node is chosen as a root, a root distribution. Depending on the form of the substitution matrices, we distinguish different evolutionary models.

Most commonly used molecular evolutionary models in phylogenetics are the so-called *continuous-time* models. In these models, the substitution events along an edge e

of a rooted phylogenetic tree occur following a continuous-time Markov process: there is a rate matrix Q that operates at intensity λ_e and for duration t_e so that the substitution matrix A^e equals $\exp(Q \cdot \lambda_e t_e)$. They are restrictive in that they assume the substitution matrices to be of exponential type and that the instantaneous mutation rate matrix Q is usually common across the tree (when this assumption holds, one talks about a *homogeneous* process). Moreover, the process is usually assumed to be also *stationary* and *time-reversible*, which imposes some other restrictions on the instantaneous rate matrix. Under the umbrella of these models fall the time-reversible models Jukes-Cantor `JC69` (Jukes and Cantor, 1969a), Kimura two-parameters, `K80` (Kimura, 1980), Kimura three-parameters, `K81` (Kimura, 1981), `HKY` (Hasegawa et al., 1985), and the General Time Reversible model, `GTR` (Tavaré, 1986).

We are interested in a broader class of evolutionary models and we model evolution using the*discrete-time* Markov processes on phylogenetic trees, i.e. we do not assume that substitution matrices are of exponential type. Among them, we find the models analogous to the continuous-time models introduced above (`JC69`*, `K80`*, `K81`*), and the more general models: `SSM` and `GMM`. In particular, these models do not impose a common instantaneous rate matrix fixed across the tree and hence different lineages in the tree are allowed to evolve at different rates (Greuel et al., 2003; Allman and Rhodes, 2004b; Semple and Steel, 2003). This modification allows to deal with the so-called *nonhomogeneous* data. In this thesis we are interested only in the discrete-time models and solve for them a number of questions that had already been addressed for the continuous-time models:

Problem 1. Provide efficient tools to select an evolutionary model that best fits the data.

Problem 2. Given a multiple sequence alignment and an evolutionary model, provide efficient tools to estimate the evolutionary parameters (both the tree topology and the substitution matrices).

Towards the goal of solving the above issues, we required a tool for generating reliable synthetic data sets.

Problem 3. Provide a method to generate data evolving along a phylogenetic tree (with given branch lengths) under a specific discrete-time evolutionary model.

The above problems have been addressed in a variety of ways for the continuous-time models (see e.g. Felsenstein (2003), Gascuel and Guindon (2007)).

The assumption that all sites in a DNA sequence are identically distributed is often too restrictive. One way of relaxing it is by considering phylogenetic mixtures. By a *phylogenetic mixture* we understand a collection of trees that altogether model the data, each being suitable for a fraction of the data set. Mixtures can include trees on different or the same tree topologies, whilst the branch lengths are allowed to vary freely. Naturally, phylogenetic mixtures best explain the heterogeneous evolutionary processes, i.e. the data comprising multiple genes or selected codon positions. Among a plethora of applications, phylogenetic mixtures are used in the orthology prediction, gene annotation, species tree reconstruction or drug target identification. In the continuous-time setting, phylogenetic mixtures are usually modeled by varying rate across site (see Semple and

Steel (2003)).

The following problem lies at the heart of applicability of phylogenetic mixture models.

Problem 4. When are the tree topologies in a phylogenetic mixture identifiable? In other words, given an evolutionary model, what conditions must be met for the existence of only one collection of tree topologies that gives rise to the observed multiple sequence alignment? Moreover, providing the tree topologies are identifiable, what conditions guarantee that the substitution parameters are identifiable (i.e. is there a single set of substitution parameters that leads to a given multiple sequence alignment)?

This problem is crucial for justifying the use of methods such as maximum likelihood. Though it has been extensively studied, at this point only a few results are known (see for instance Allman and Rhodes (2006a), Allman et al. (2010), Stefanovic and Vigoda (2007), Rhodes and Sullivant (2011),Chai and Housworth (2011)).

Problem 1 can be rephrased in reference to phylogenetic mixtures. Indeed, when choosing an evolutionary model that best fits the given data, a phylogenetic tree is unknown. Therefore, the interest lies in the evolutionary model that best fits the data under the assumption of the data had evolving along any tree or a mixture of trees for the model considered. The problem of choosing the most suitable model for the given data is usually a heuristic choice, and currently there exist no methods that do not rely on a circular argument of an estimated input tree (cf. Posada and Crandall (2001)).

This leads to the following question:

Problem 5. Is there a way to characterize distributions that arise from phylogenetic mixtures under a given evolutionary model? If so, can we use it as a tool for model selection?

In this thesis we approach the above problems from the standpoint of algebraic statistics and in most part they are solved for the `JC69`*, `K80`*, `K81`*, `SSM` and `GMM` models. The solution to the problems 1, 4, and 5 requires a deep mathematical study of these models. These are instances of the so-called *equivariant models*, whose symmetries in the transition matrices give rise to appealing properties of the distributions of the DNA sequences in a MSA that evolved under these models. With the purpose of studying the properties of these models, we use the techniques from algebraic geometry and group theory.

Indeed, it is well known that the expected probabilities of nucleotides observed at the leaves of a phylogenetic tree satisfy a given collection of equalities if the tree evolved under certain models (see for instance (Felsenstein, 2003, p.375)). It was already pointed out by Fu and Li (1992a), Steel et al. (1992) or Felsenstein (2003), that these equalities (referred to as *linear invariants*) could potentially be used to test the evolutionary model the data came from. How can one guarantee that there are no more equalities to be used? We employ tools from algebraic geometry to answer these questions and to address the identifiability issue for phylogenetic mixtures. Furthermore, we use statistical techniques to provide an efficient model selection algorithm using algebraic model invariants.

We solve problem 2 by adapting the well-known Expectation-Maximization algorithm to our setting. Tree inference is beyond the scope of this thesis. In order to solve problem 3 we use basic algebraic techniques.

To summarize, in this thesis we achieved the following goals:

- Provide algorithms for generating *any* substitution matrix with a given branch length under the JC69*, K80*, K81* and the SSM models (and some matrices for the GMM model);
- Implement these algorithms in the package GenNon-h, which generates multiple sequence alignments evolving along a tree under any of these models and any number of taxa;
- Implement the Expectation-Maximization algorithm to provide the maximum likelihood estimates of the entries of the substitution matrices and the root distribution given a multiple sequence alignment on any number of taxa, a tree topology and an evolutionary models above. We implement it in a package called Empar, which in addition performs statistical tests for the parameter estimates;
- For the trees on 4 and 6 taxa, perform an in-depth study of the performance of Empar and its dependence on factors such as model complexity, size of the tree, positioning of the branches, data and total tree lengths;
- Characterize the distributions arising from phylogenetic mixtures under the JC69*, K80*, K81* and SSM models (see Theorems 6.7 and 6.11);
- Use the above characterization (in a maximum likelihood framework) and the Akaike's information criterion in model selection, and implement it as SPIn for any trees and any number of taxa;
- Test the successfully performance of SPIn on simulated and real data and compare it to other existing methods;
- Provide an upper bound on the number of tree topologies that are identifiable for phylogenetic mixtures under the models considered here;
- Use the above methods to characterize the evolutionary patterns of the regions annotated in the *GENCODE* project.

The algorithms mentioned above have been implemented in C++ under the names: GenNon-h, Empar, and SPIn. They are freely available on the following pages:

http://genome.crg.es/cgi-bin/phylo_mod_sel/AlgGenNonH.pl

http://genome.crg.es/cgi-bin/phylo_mod_sel/AlgEmpar.pl

http://genome.crg.es/cgi-bin/phylo_mod_sel/AlgModelSelection.pl.

Some of the results of this thesis have been published in the paper Kedzierska et al. (2012) and in the preprint Casanellas et al. (2011). One additional article is under revision and two in preparation.

The thesis is structured as follows:

Part I provides the required biological background, both from the genomic and phylogenetic perspective. In this part the reader will find information on the evolutionary models employed in this work. We then present and shortly discuss a motivating study undertook at the conception of the project. This is a case study across-species conservation of motifs involved in the regulation of splicing. We developed an approach that estimates a conservation of sequence motifs with a sitewise precision by incorporating phylogenetic information, and is able to detect even weak selective constrains. As expected, we found that distinct varying levels of positive selection can be found even within short sequences.

These observations give a hint that phylogenetic mixtures are possibly an underestimated tool in phylogenetics.

Part II presents all our theoretical results. To start, in chapter 3 we derive algorithms for generating substitution matrices with a given branch length under a selected equivariant model considered in this work. Next, chapter 4 contains the details of the Expectation-Maximization algorithm for parameter estimation. In section 5.1 the reader can find the background required for understanding this thesis. We follow by an introduction to the concepts of algebraic evolutionary models and invariants in sections 5.2 and 5.3, while background on group theory is contained in section 5.4. Examples and a detailed study of certain equivariant models from the perspective of group theory is given in section 5.5. In section 6.3 we prove that the space of distributions arising from phylogenetic mixtures evolving under an evolutionary model is determined by a linear space. By exhaustively studying the group of symmetries of these models, we give an easy and combinatorial way of determining the equations of this linear space for the equivariant models considered in this work. These linear equations are at the foundation of the method we developed for model selection. As a last theoretical component of the thesis, chapter 6 is dedicated to the study of phylogenetic mixtures and their identifiability.

Part III contains the details on the implementations and tests of the methods developed in the course of this work. Section 7 is dedicated to `GenNon-h`, section 8 to `Empar` and section 9 to `SPIn`. We then present results on the applications of the two latter methods to the data from the *ENCODE* project in chapter 10. Lastly, chapter 11 contains the possibilities for future work.

Part I

Biological motivation

Chapter 1
Biological preliminaries

1.1 Central dogma of molecular biology

It is currently accepted that life on Earth is approximately 3.5 billion years old, dividing all living organisms into 3 domains (Eukaryota, Bacteria, Archarea) and 5 kingdoms: Monera (bacteria), Protista, Fungi, Plantae, and Animalia. Darwin's theory of evolution by natural selection (Darwin, 1929) describes the process of evolution of organisms and speciation. Darwinian evolution states that at all points in time living creature went through a process of variation. The individuals most successful to survive in a given environment are those who reproduce most successfully and pass on greater number of their traits to their offspring. This theory gave rise to the hypothesis that the diversity of life forms on Earth comes from the divergence of one common ancestral unicellular organism.

Evolution can be extended to the levels of *DNA (deoxyribonucleic acid)* and proteins as all living organisms can replicate by means of DNA and are able to convert the information stored in DNA into cell-building products.

DNA structure was deciphered by James D. Watson and Francis Crick in 1953 Crick and Watson (1953). Watson and Crick discovered the DNA as the molecular basis of heredity. DNA molecule is composed of two anti-parallel strands of nucleotides, which form a double helix. The nitrogenous base is directed towards the axis of the structure, while the two backbones are composed of sugar and phosphate alternating units. Accordingly to the four bases, the nucleotides are: adenine (`A`), thymine (`T`), guanine (`G`) and cytosine (`C`). We will refer to `A`, `C`, `G`, `T` as bases or nucleotides interchangeably.

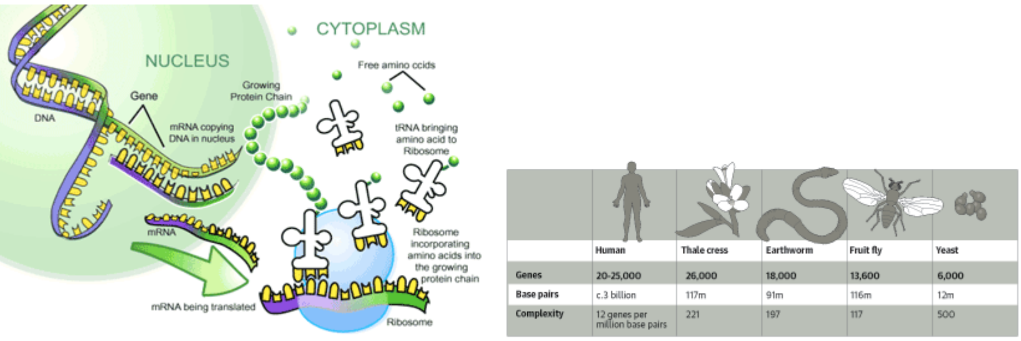
Figure 1.1: DNA

The complementary bases, A − T and C − G, joined by hydrogen bonds are referred to as *base pairs, bp*. The hydrogen bonded base pairs form the core of the molecule. The base pairs stack on top of one another and parallel to other pairs each in a spacing of 3.4 angstroms. The convention is to impose a direction on the chain: from the *5' end* with an exposed phosphate group (positioned as the left end), to the *3' end* with an exposed ribose group (the right end).

The size of the human genome is around 3 billion bp. The sizes of genomes and numbers of genes they contain vary between the species. Positive correlation between the number of genes and the complexity of an organism has many exceptions. For example, the genome of the brown mountain grasshopper (*Podisma pedestris*) is seven times larger and the genome of onion is six times larger then the genome of humans (Bensasson et al., 2001; Jakse et al., 2008). An average length of human genes is around 3.000*bp* (see Fig. 1.2). Diversity of the human genome and other complex organisms lies in the use of alternative splicing (see below; Xing and Lee (2006)).

The discovery of the DNA allows the evolution to be explained in a new way.

The central dogma of molecular biology describes the process of protein synthesis. Based on the finding that DNA and RNA are build of a similar and the specific chemical pairing of nucleotides occurs, Crick suggested that DNA can be used as a template for RNA synthesis. The two major steps include the processes of *transcription* and *translation*. In brief, DNA makes RNA via what is called transcription and via translation RNA makes protein. The process is illustrated in Figure 1.2. DNA is replicated and transcribed to RNA, which codes for one or more genes (*transcription unit*). In this step, RNA polymerase catalyzes the formation of the primary gene transcript, the *precursor mRNA (pre-mRNA)* molecule. Pre-mRNA is additionally modified by *splicing machinery*: stretches of sequences that code for a protein, *exons*, are kept and *introns*, which are the non-coding parts, are removed (see Fig. 1.1). If this RNA is a blueprint for protein, the RNA becomes the *messenger RNA* (mRNA). In eukaryotic cells the mRNA is spliced and migrates from the nucleus to the cytoplasm. Final step is translation– mRNA encodes the information, which is "read off" by ribosomes and used for protein synthesis. Genetic information in DNA and RNA is coded in triplets of nucleotides (*codons*). Except for the start and stop codons occurring at the two ends


CYTOPLASM
NUCLEUS
Free amino ccids
Growing Protein Chain
Gene
DNA
mRNA copying DNA in nucleus
tRNA bringing amino acid to Ribosome
mRNA
Ribosome incorporating amino acids into the growing protein chain
mRNA being translated
Ribosome

	Human	Thale cress	Earthworm	Fruit fly	Yeast
Genes	20-25,000	26,000	18,000	13,600	6,000
Base pairs	c.3 billion	117m	91m	116m	12m
Complexity	12 genes per million base pairs	221	197	117	500

Figure 1.2: Central dogma of molecular biology (adapted from `http://www.scq.ubc.ca/`)

of the transcript and denoting the beginning and termination for the protein synthesis, each codon contains a specific amino acid information. The translation machinery is located within a *ribosome*– a specialized organelle containing *ribosomal RNA* (*rRNA*) and *transfer RNA* (*tRNA*). In eukaryotes ribosomes are located in cytoplasm and is composed of two units, the small and large one, which travel separately, but enclose around the mRNA to start the translation process. Based on the complementary base-pairing, on one side the tRNA molecules read the triplet code in the mRNA and attach to a specific amino acid on the other. rRNA catalyzes the process of attaching of newly created amino acid to the growing protein chain.

Proteins are involved in all biological structural and enzymatic activities. The whole process of "manufacturing" protein from a given gene is called *gene expression*.

Splicing is a process that takes place at the level of preRNA. It is known that a about $\sim 94\%$ of human genes code for more than one protein (Pan et al., 2008). The two ends of the coding parts of the transcript are marked by the so-called *splice sites*, which are consensus sequences of nucleotide bases denoting the start and end of a gene (see Fig. 1.1). The process of splicing can be *constitutive* or *alternative*.

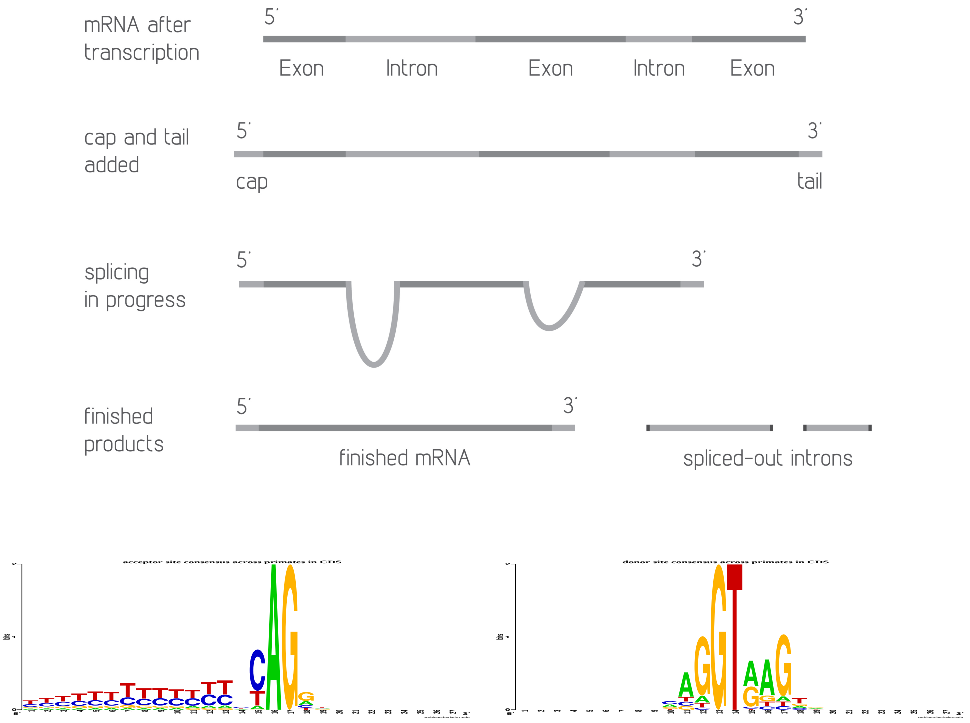


Figure 1.3: *Top*: illustration of splicing; *bottom*: consensus sequence of the 3' and 5' splice site of the transcript.

Constitutive splicing removes all introns in the pre-mRNA and re-connects all exons into the final transcript. Alternative splicing connects the exons in a variety of ways, which in turn creates distinct transcripts and leads to different protein isoforms. Exonic and intronic *Splicing Enhancers* (*ESEs, ISEs*) and *Silencers* (*ESSs, ISSs*) are factors playing a crucial role in this differentiation process. These short sequences ($\sim$ 8 nucleotide long) either enhance or silence the neighbouring splice sites and their recognition by the splicing machinery.

1.2 Assembly and annotation of the human genome

Celera Corporation (Venter et al., 2001) and the *International Human Genome Sequence Consortium* (Lander et al., 2001) independently completed sequencing of the first draft of the human genome. These initial drafts were further refined to create a consensus and a high quality standard version of the human genome sequence. This is the task of the *Genome Reference Consortium* (*GRC*) (`http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/`). *GRCh37* (*Human Reference, hg19*, Fujita et al. (2010)) is the last release of the human genome assembly from February 2009. Genome resources for human and other species are hosted at the *NCBI* website. The GRCh37 build assembly can be found on the NCBI Build 37.1 Statistics page: `http://www.ncbi.nlm.nih.gov/mapview/stats/BuildStats.cgi?taxid=9606&build=37n&ver=1`.

A pilot phase of the project called *The Encyclopedia Of DNA Elements, ENCODE,* was launched in 2003 (TheENCODEProjectConsortium (2007), `http://genome.ucsc.edu/ENCODE/`, (`http://encodeproject.org`). During its duration of 4 years, its goal was to identify functional elements in 1% of the human genome sequence. Focus was also placed on developing methodologies for analyzing, storing and sharing data. In 2007 the study was extended to the entire genome (TheENCODEProjectConsortium, 2011).

The *GENCODE* Project– *The Encyclopedia of genes and genes variants*– is a subproject of *ENCODE* that started with the objective of identification of all splice variants of protein-coding genes within the 1% *ENCODE* regions in the human genome (Harrow et al., 2006). During the extended phase of *ENCODE*, *GENCODE* is also annotating splicing variants of long non coding RNAs, and small RNAs.

1.3 Multiple sequence alignments

DNA sequence refers to the sequence of nucleotides in a single strand of a DNA molecule. Its length is measured in nucleotides (nt) or, equivalently, in base pairs (bp).

Sequence alignment is a way of arranging biological sequences (DNA, RNA, or protein) to identify homologous regions (similarity of aligned sequences poses a hypothesis about their inheritance from a common ancestor). Aligned sequences are typically presented as rows in an array whose columns are formed by characters that have (presumably) evolved from the same character on the common ancestral sequence. These so so-called *multiple sequence alignment* (MSA) allow for simultaneous comparison of several sequences. MSA serves the purpose of identifying regions that may be a consequence of functional, structural, or evolutionary relationships between the sequences. MSA are used in many contexts, including phylogenetic analysis, sequence-pattern recognition or identification of functional elements.

Fasta format became a standard for representing information contained in the MSA. Every record starts with a symbol > followed by sequence identifier (protein or gene name and source organism) and by a nucleotide or protein sequence split in blocks, e.g.

```
>hg19
CCCTTTGTACCAGTTGTAGCCATAAAGATTCTGGGACTCATTATGGACTACTAGAAGGACCTCCTT
CCCTTCTGCGACATTGAACGGCACGACATCAATATTGGTCTGGGCACTGTT
>mm9
TCCCTTGTACCAGGCAAAGGCTCCAAGCGCCAGGGGCAGATTGTGAACAAGTAGAAGAACATTGTT
GTCTTCAGCAACCTGGGGCGGCACAGCCTCAATGGTGACTTCAGCAGTGGT
>rn4
TCCCTTGTACCAGTAAAAGACTTGGAACTCCTGCGGCAGATTGTGAGCGAGTAAAAGAACGCTCTT
CTCCTCAACAACGTTGGGTGGCACAGCGTCTACGGTGACTTGGGCAGTGGT
```

is a fragment of a MSA of the DNA of human, mouse and rat.

Aligning sequences can be performed for both DNA and protein sequences. Pairwise alignment is *global*, when it is performed on the full-length sequences. Greedy algorithms or variants of the dynamic Needleman-Wunsch algorithm (Needleman and Christian, 1970) lie at the base of development of the software created for performing pairwise sequence alignments e.g. *BLAST* Altschul et al. (1990), *FASTA* Pearson and Lipman (1988), *Align* at *EMBOSS*, `http://www.ebi.ac.uk/Tools/psa/`). Some of the most popular DNA alignment programs include *ClustalW(2) (Thompson et al., 1994), T-coffee (Notredame et al., 2000), MUSCLE (Edgar, 2004), MAVID (Bray and Pachter, 2003).*

Alignment programs introduce gaps (denoted by '-') in sequences relative to others in order to provide better quality alignments. In this work, we will consider gap-free alignments.

1.4 Phylogenetics: Markov models of evolution

Systematics is the field of biology which examines the natural variation and relationships of organisms. Taxonomy is one of its branches and deals with the nomenclature, identification and classification of organisms. Often the terms taxonomy and systematics are used interchangeably. The so-called *operational taxonomic units* (OTUs) represent the organisms alive today (plants, animals, microorganisms). The relationships between OTUs, which are a result of this classification are represented on graphical structure of trees. In this section we get acquainted with the notion of phylogenetic trees and evolutionary models.

1.4.1 Phylogenetic trees

Definition 1.1. A *tree* $\mathcal{T}$ is a connected acyclic graph, which consists of a set of vertices and edges that connect them.

We distinguish two types of vertices: leaves, which are the terminal nodes, and the interior nodes, $Int(\mathcal{T})$. We denote by $\mathtt{L}(\mathcal{T})$ the set of leaves, by $E(\mathcal{T})$ the set of edges and by $N(\mathcal{T})$ the set of all nodes of $\mathcal{T}$. The elements of $\mathtt{L}(\mathcal{T})$ correspond to OTUs. Most often, the information contained in the leaves comes from either the DNA or protein sequences within one or more organisms.

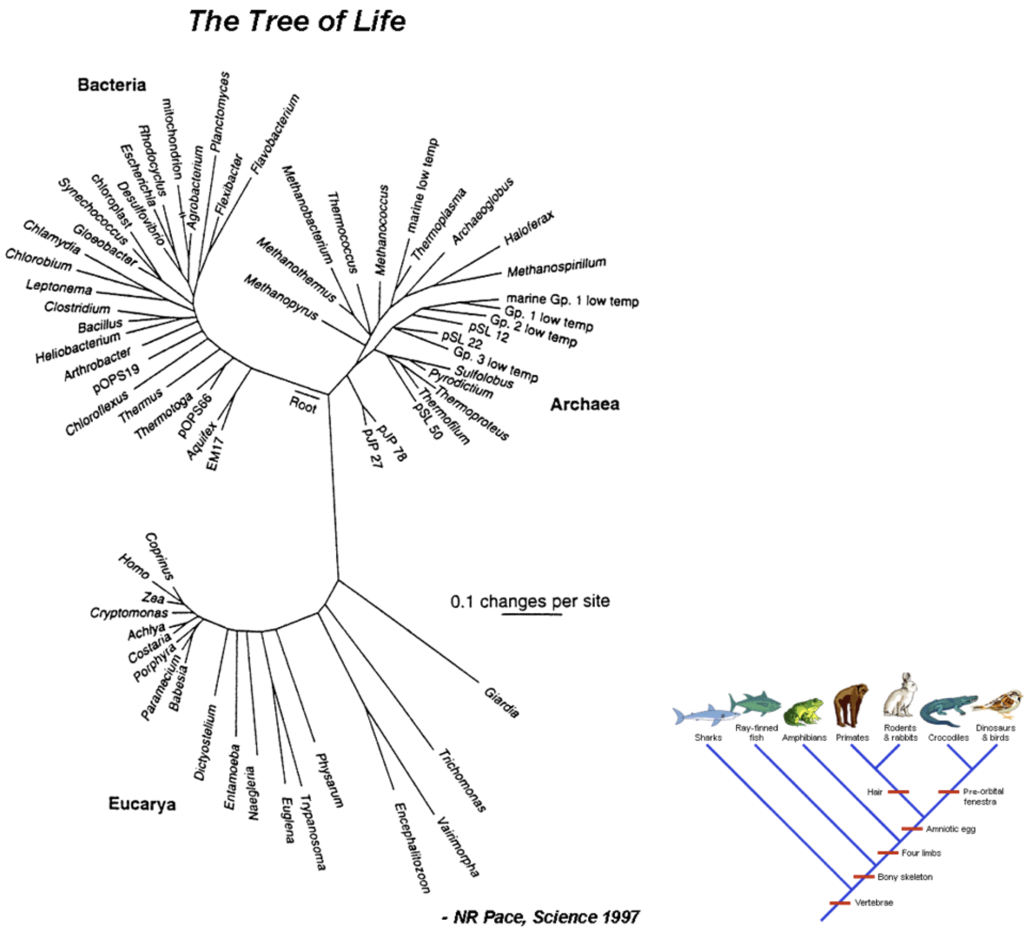


Figure 1.4: Tree of life and one of its fragments (Adapted from evolution.berkeley.edu)

We will also write $e = (e0, e1)$ for an edge $e \in E(\mathcal{T})$, where e_0 and e_1 are two ends of e.

Definition 1.2. A *phylogenetic tree* is a triplet $(\mathcal{T}, \rho, \{v_1, \ldots, v_n\})$ where $\mathcal{T}$ is a tree with n leaves, $\{v_1, \ldots, v_n\}$ is a set of different sequences, and $\rho : \{v_1, \ldots, v_n\} \longrightarrow \mathtt{L}(\mathcal{T})$ is a bijection.

Definition 1.3. A *rooted phylogenetic tree* is a tree where a distinguished interior vertex is selected to be the *root* $\mathbf{r}$ (see Fig. 1.6(b)). The root is usually imposed and represents the last common ancestor of the set of observed sequences. It induces an orientation on the edges of $\mathcal{T}$. The leaves represent information about current sequences (present) and the interior nodes represent ancestral sequences (past), thus a tree records the ancestral relationships among the current species. The vertex adjacent to an edge e that lies closer to the root is called an *ancestor*, while the other end is a *descendant*. The *degree* of a vertex is the number of its outgoing edges. A rooted tree is *binary* if $\mathbf{r}$ has degree two and the remaining nodes are of degree three. An unrooted tree is called *trivalent* if all its interior nodes have a degree three. A *star tree*, also referred to as a *claw* tree, is a tree with only one interior node. In the n-taxon star tree the root has degree n.

Therefore, there are two kinds of information encoded in a phylogenetic tree:

1. the structure of the labeled graph, called the *tree topology* (see Def. 1.5), which represents evolutionary relationships among a set of sequences that are believed to have a common ancestor,

2. the *length* of the edges, so called *branch lengths* measuring evolutionary time, i.e. the amount of nucleotide changes accumulated between the e_0 and e_1 ends of e.

An estimated number of species inhabiting Earth is 5 to 100 millions out of which only 1-2 million are classified and named. The *tree of life* is a graphical representation of the relationships between the forms of life (see Fig. 1.4.1) and their evolution from a common ancestor. The Tree of Life organizes the knowledge about the history of lineages on the axes of time and is based on the assumption that species arise from the previous ones by descent and that all organisms are connected via passage of genes. Internal nodes correspond to division or speciation events (when new biological species arise) leading to independently evolving lineages. Under the framework of the *Tree of Life Project* (National Science Foundation, `http://www.phylo.org/atol/`), it undergoes continuous updated as the new information and discoveries become available. Any phylogenetic tree is therefore a subtree of the Tree of Life.

Remark 1.4. Nucleotide changes alter the genetic information carried by a given gene. Substitutions, deletions, insertions and mutations are changes to the genetic sequence, thus shape the composition of the genomes. Substitution is a change of one nucleotide to another that become fixed within population ("tolerated" by evolution in at least their last common ancestor). Mutations happen due to mistakes in DNA replication or repair. They refer to the alterations at both large and small levels, both gross chromosome or small point mutations. The latter can involve a change at a single position in a nucleotide sequence. The changes can be caused by a variety of external and internal mutagenic agents (i.e. chemical mutagens, radiation, sunlight, spotaneous changes of isomers) and it can be deletarious, advantageous or neutral. Sometimes mutations and substitutions are used interchangeably. In this work we will focus on substitutions. *Synonymous substitutions* in the protein coding exons are substitutons that do not modify the resulting amino acid. Otherwise, they are called *non-synonymous.*

Definition 1.5. Let $(\mathcal{T}_1, \rho_1, \{\mathtt{u}_1, \ldots, \mathtt{u}_n\})$ and $(\mathcal{T}_2, \rho_2, \{\mathtt{u}_1, \ldots, \mathtt{u}_n\})$ be two phylogenetic trees with the same set of leaves. We say that they have the same *tree topology* if there exists an homeomorphism $f : \mathcal{T}_1 \longrightarrow \mathcal{T}_2$ such that

$$f(\rho_1(\mathtt{u}_i)) = \rho_2(\mathtt{u}_i) \quad \forall i = 1, \ldots, n.$$

If $\mathcal{T}_1$ and $\mathcal{T}_2$ are rooted and $\mathbf{r}_1$, $\mathbf{r}_2$ are their respective roots, we will also impose that $f(\mathbf{r}_1) = \mathbf{r}_2$.

Another commonly used term in phylogenetics is a *clade* and an *outgroup*. A phylogenetic tree is composed of clades that can be thought of different evolutionary lines. A clade is a grouping that incudes an ancestor and all its descendants. Clades can consist of a few or a large number of species. They form a nested hierarchy: smaller clades are included in the bigger ones and cladistics is a method that deals with such classification. An outgroup, on the other hand, is a taxon separated from the rest of taxa by a larger evolutionary distance and stems from last (hypothesized) common ancestor of the organisms under study. Based on the assumption that species evolve by descent with modification, it is often used to determine the shared derived characteristics of sequences under study and is very useful for phylogenetic tree reconstrucion. For instance, in the fragment of the phylogenetic tree classifying the *Drosophila* genus depicted in Figure 1.5, the representant of the *Hawaiian Drosophila* can be taken as an outgroup to the subgenus *Sophophora*.

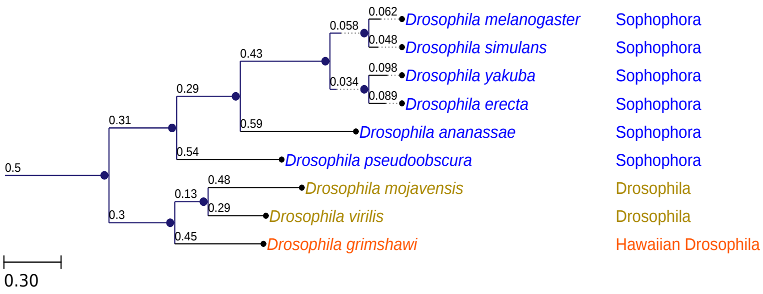


Figure 1.5: Phylogenetic tree of 9-taxon drosophila (Pollard et al., 2006a; Clark et al., 2007)

The number of edges of trivalent trees on n leaves is $2n-3$. The number of distinct tree topologies of trivalent unrooted trees on n leaves is $(2n-5)!!$, while the number of rooted tree topologies is $(2n-3)!!$. These numbers show that the task of finding the correct underlying tree or deciding on the most suitable model is nontrivial– more than an exponential increase in the number of leaves presents a challenge, both conceptual and computational.

1.4.2 Hidden Markov processes on trees

We adopt a probabilistic view on modeling evolution. Evolution is assumed to be a stochastic process, in which nucleotides evolve over time according to certain probabilities.

Changes that can be observed between two DNA sequences are described as substitutions, insertions or deletions. In the two latter cases, a nucleotide is inserted or deleted from a given position as compared with the other sequence. In most commonly used evolutionary models insertions and deletions are not considered and incorporating them would highly increase the complexity of the model. Throughout this work we focus solely on the substitutions occuring along the evolutionary process (no insertions nor deletions).

Let $\mathcal{T}$ be a rooted phylogenetic tree on n taxa labeled as $\{1, \ldots, n\}$. We adopt the orientation from the root to the leaves of $\mathcal{T}$. We assume that the sites in the alignment are independent and identically distributed. That is, the states at each position in the sequence evolve independently of the other nucleotides and according to the same evolutionary process. It is disputable whether this assumption is realistic, however, models are mere simplification of the evolutionary process, and certain assumptions allow for more convenient inferential frameworks. We associate a discrete random variable to each node of $\mathcal{T}$ with k possible states and we assume a fixed order on the k states. Usually, k is taken to be 4, representing the four main bases in DNA, in which case the states are Adenine, Cytosine, Guanine, Thymine denoted in this order as $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. The random variables at the leaves are observed, while the random variables at the interior nodes are hidden.

Let $\pi = (\pi_1, \ldots, \pi_k)$ be the distribution of the k states at the root of $\mathcal{T}$. It has $(k-1)$ degrees of freedom due to the constraint $\sum_{i=1}^{k} \pi_i = 1$. It is easy to see that the maximum likelihood estimates, $\hat{\pi}$, of π are the relative frequencies of each nucleotide in the ancestral sequence assigned to $\mathbf{r}$. For example, if the sequence is TCAACTGATC with the states $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, then we have that $\hat{\pi}_{\mathtt{A}} = \frac{3}{10} = \hat{\pi}_{\mathtt{C}} = \hat{\pi}_{\mathtt{T}}$, $\hat{\pi}_{\mathtt{G}} = \frac{1}{10}$.

Now, to each edge e of $\mathcal{T}$ we associate a $k \times k$ transition matrix A^e whose entries are indeterminates representing the probabilities of transition between the two ends of e. Markov assumption means that the current state of the process is dependent only on the most immediate ancestral state. That is to say, the evolutionary process at two bifurcating branches are *independent given* the common node. Let us recall that two random variables A and B are *independent given* a third random variable C if $P(A, B \mid C) = P(A \mid C)P(B \mid C)$.

More formally, a *transition (substitution) matrix* for a Markov model on $k = 4$ states, $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, is defined as

$$A^e = \begin{pmatrix} P_{\mathtt{A}|\mathtt{A}} & P_{\mathtt{C}|\mathtt{A}} & P_{\mathtt{G}|\mathtt{A}} & P_{\mathtt{T}|\mathtt{A}} \\ P_{\mathtt{A}|\mathtt{C}} & P_{\mathtt{C}|\mathtt{C}} & P_{\mathtt{G}|\mathtt{C}} & P_{\mathtt{T}|\mathtt{C}} \\ P_{\mathtt{A}|\mathtt{G}} & P_{\mathtt{C}|\mathtt{G}} & P_{\mathtt{G}|\mathtt{G}} & P_{\mathtt{T}|\mathtt{G}} \\ P_{\mathtt{A}|\mathtt{T}} & P_{\mathtt{C}|\mathtt{T}} & P_{\mathtt{G}|\mathtt{T}} & P_{\mathtt{T}|\mathtt{T}} \end{pmatrix},$$

where the conditional probability $P_{i|j}$ denotes a change (substitution) of the nucleotide i at the node e_0 to the nucleotide j at e_1.

Definition 1.6. A square matrix A^e is called a *stochastic matrix* if it has row sums equal to 1 and nonnegative real entries. It is called *strictly stochastic* if moreover all its entries are strictly positive.

We will denote the $P_{j|i}$ entry of A^e by A^e_{ij}. An evolutionary Markov process is therefore characterized by the set of parameters of the root and substitution matrices (see Fig. 1.6(b)). If no other restrictions on model parameters are present, the number of degrees of freedom is $3 + 12|E(\mathcal{T})|$. According to the shape of the transition matrices and the root distribution we have different evolutionary models as we will see in the examples later in this section. The tree topology and the entries of the transition matrices are the parameters of a model and the goal of phylogenetic inference is to estimate them from the observed data of of DNA sequences.

Let us recall a few facts about the Markov matrices. Markov assumption means that given two evolutionary processes: (π^e, A^e) from e_0 to e_1 and $(\pi^{e'}, A^{e'})$, from $e_0^{'}$ to $e_1^{'}$ such that $e_0^{'} = e_1$, we have that $\pi^{e'} = \pi^e A_1 = \pi^e A^e A^{e'}$. If A^e and $A^{e'}$ are Markov matrices of the same size, then $A^e A^{e'}$ is also a Markov matrix. The condition of row sums equal to one is equivalent to stating that $A\mathbf{1} = \mathbf{1}$, where $\mathbf{1} = [1, \dots, 1]^t$. Therefore, 1 is an eigenvalue of any substitution matrix.

Theorem 1.7 (Perron–Frobenius, Chang et al. (2008))**.** *Let A be a Markov matrix. Then every eigenvalue λ of A satisfies $|\lambda| \leqslant 1$. Moreover, if A has positive entries, then 1 is a simple eigenvalue (has multiplicity 1) and $|\lambda| < 1$ for any other eigenvalue λ; in addition,* $\dim \operatorname{Ker}(A - \mathrm{id}) = 1$.

The above theorem ensures that the limit $\lim_{m\to\infty} A^m$ exists.

Given a vector π, we denote by D_π the $k \times k$ diagonal matrix with the vector π on its diagonal.

Definition 1.8. A Markov process on a rooted phylogenetic tree $\mathcal{T}$ is *stationary* with an *equilibrium* vector π if $\pi = \pi A^e$ for all $e \in E(\mathcal{T})$. If the equilibrium distribution exists, it is unique and $\lim_{m\to\infty}(A^e)^m = \pi$. A stationary process is said to be *time-reversible* if $D_\pi A^e = (A^e)^T D_\pi$ for all $e \in E(\mathcal{T})$.

If a model is stationary with an equilibrium vector π, then the distribution at the root of $\mathcal{T}$ is usually taken to be equal to π as well.

A way of specifying the evolutionary model is to assume a *continuous-time homogeneous* Markov process along each edge. Usually, the instantaneous rate of substitutions is common in the entire tree and is recorded in a rate matrix Q. every edge of the $\mathcal{T}$. The rate matrix is set to have negative diagonal elements and the diagonal entries chosen such that the rows sum to 0: $Q\mathbf{1} = 0$. If the root distribution is taken to be the eigenvector corresponding to the eigenvalue of 0, then the process is stationary. Following on Felsenstein (2003) Q is often assumed to be time-reversible (see Def. 1.8).

If t_e is the length of the branch in $\mathcal{T}$, then the substitution matrix on the edge e is given by the set of differential equations $A^e(t)' = QA^e(t)$ with $A^e(0) = id$. The solution to these equations is given by the matrix exponential $A^e = \exp(t_e Q)$.

In contrast to these continuous-time Markov processes, the Markov models introduced at the beginning of this section can be thought of as discrete-time Markov processes. Discrete-time formulation of the models allows for more flexible framework of distinct rate matrices for different lineages. Indeed, even if the substitution matrices A^e are of exponential type, their logarithms do not need to be proportional. Moreover, not all substitution matrices can be represented as an exponential of a real matrix (see Remark 3.10).

In phylogenetic inference, the tree topology $\mathcal{T}$ is a discrete parameter. Given a model $\mathcal{M}$ and a tree $\mathcal{T}$, the continuous parameters are the root distribution π and a set of substitution matrices $(A^e)_{e\in E(\mathcal{T})}$ that satisfy model requirements. Let $p_{\mathcal{T}}^{\mathcal{M}}(\pi, (A^e)_{e\in E(\mathcal{T})})$

be the vector of joint probability distribution of the states observed at the leaves of $\mathcal{T}$ under the Markov process. Its entries are the 4^n probabilities $p^{\mathcal{M}}_{\mathcal{T},\mathtt{x}_1\ldots\mathtt{x}_n}(\pi,(A^e)_{e\in E(\mathcal{T})})$ of observing each nucleotide pattern $(\mathtt{x}_1\ldots\mathtt{x}_n)$ at the leaves of $\mathcal{T}$ as given by the parameters $\{\pi,(A^e)_{e\in E(\mathcal{T})}\}$. For simplicty of exposition, we will denote it as $p_{\mathtt{x}_1\ldots\mathtt{x}_n}$ whenever $\mathcal{M}$ and $\mathcal{T}$ are clear from the context. According the Markov process on the tree $\mathcal{T}$ we

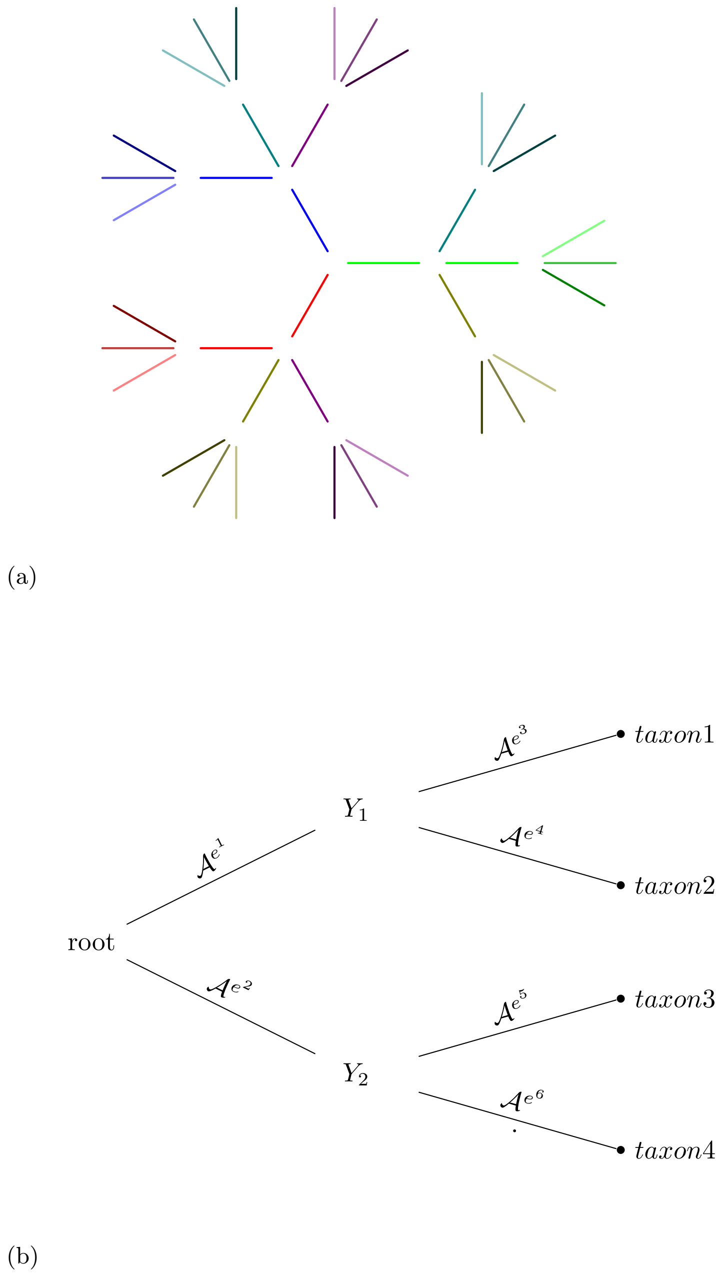


Figure 1.6: Examples of phylogenetic trees: a) A circular (unrooted) tree with vertices of degree 3 and 4;b) A bifurcated 4-taxon tree with vertices $\{root, Y_1, Y_2\}$. The transition matrices, A^{e^i}, of the 6 labeled branches, e^i together with a root distribution define an evolutionary model.

have

$$p_{\mathtt{x}_1\ldots\mathtt{x}_n} = \sum_{\mathtt{x}_v \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}, v\in Int(\mathcal{T})} \pi_{\mathtt{x}_r} \prod_{v\in N(\mathcal{T})\setminus\{r\}} A^{e_{an(v)},e_v}_{x_{an(v)},x_v} \tag{1.1}$$

$\mathtt{x}_u$ denotes the state at the vertex u and $an(v)$ is the parent node of v. If $v = i$ is a leaf node, then $\mathtt{x}_v = \mathtt{x}_i$.

Example 1.9. Consider a claw tree $\mathcal{T}$ with 3 labeled leaves $\{X_1, X_2, X_3\}$ and a general model on the four states $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Following a hidden Markov process on $\mathcal{T}$ we can write explicit formulas for the joint probability:

$$p_{\mathtt{x}_1\mathtt{x}_2\mathtt{x}_3} = \sum_{b\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_b A^{r,v_1}_{b,\mathtt{x}_1} A^{r,v_2}_{b,\mathtt{x}_2} A^{r,v_3}_{b,\mathtt{x}_3},$$

where $A^{r,v_1}_{b,\mathtt{x}_1}$ corresponds to the entry in the row labeled by b and the column by $\mathtt{x}_1$ of the branch first branch. We can write the 4^n joint probabilities:

$$\begin{aligned}
p_{\mathtt{AAA}} &= \sum_{b\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_b A^{r,v_1}_{b,\mathtt{A}} A^{r,v_2}_{b,\mathtt{A}} A^{r,v_3}_{b,\mathtt{A}} \\
p_{\mathtt{AAC}} &= \sum_{b\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_b A^{r,v_1}_{b,\mathtt{A}} A^{r,v_2}_{b,\mathtt{A}} A^{r,v_3}_{b,\mathtt{C}} \\
&\vdots \\
p_{\mathtt{TTG}} &= \sum_{b\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_b A^{r,v_1}_{b,\mathtt{T}} A^{r,v_2}_{b,\mathtt{T}} A^{r,v_3}_{b,\mathtt{G}}. \\
p_{\mathtt{TTT}} &= \sum_{b\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_b A^{r,v_1}_{b,\mathtt{T}} A^{r,v_2}_{b,\mathtt{T}} A^{r,v_3}_{b,\mathtt{T}}.
\end{aligned}$$

Some of the most established discrete-time evolutionary models were first introduced by Allman and Rhodes (2007) .

Definition 1.10. The General Markov model ($\mathtt{GMM}$, Allman and Rhodes (2003); Steel et al. (1994)), the most general of the discrete-time models, has the Markov transition matrices of the form:

$$\begin{pmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \\ m & n & o & p \end{pmatrix}, \quad \text{with} \quad \begin{aligned} a+b+c+d &= 1, \\ e+f+g+h &= 1, \\ i+j+k+l &= 1, \\ m+n+o+p &= 1. \end{aligned}$$

The only restrictions in the paramaters are the stochastic conditions of the matrices (12 free parameters) and the root distribution (3 free parameters).

Consequently, one can define its submodels by imposing additional restrictions on the model parameters.

Definition 1.11. Continuous-time Jukes-Cantor model was introduced by Jukes and Cantor (1969b) and is the simplest of the possible models. It has only one free parameter representing a substitution to a distinct nucleotide base. The transition matrices for the discrete-time Jukes-Cantor model, `JC69`*, are of type:

$$\begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix}, \quad \text{with } a + 3b = 1.$$

Definition 1.12. The model introduced in Kimura (1980) has equal base frequencies and 2 free parameters corresponding to different rates of transition (interchanges of purines, A ↔ G, or pyrimidines C ↔ T) and transversions (purines-pyrimidine changes). The transition matrix for the discrete-time version of this model, `K80`*, is:

$$\begin{pmatrix} a & b & c & b \\ b & a & b & c \\ c & b & a & b \\ b & c & b & a \end{pmatrix}, \quad \text{with } a + 2b + c = 1.$$

Definition 1.13. The Kimura 3-parameter model was introduced in Kimura (1981) as an extension to the previously described model with an additional parameter corresponding to different rates of transversions. The transition matrices for its discrete-time version are of type:

$$\begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix}, \quad \text{with } a + b + c + d = 1.$$

The `JC69`*, `K80`*, `K81`* models have a stationary uniform distribution $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. This is taken to be the root distribution as well.

Notation 1.14. As seen on the above definitions, we use the symbol $(^*)$ to emphasize the nonhomogeneous nature of these models and to distinguish them from their respective continuous-time correspondents. Therefore, we write `JC69`* for the discrete-time Jukes-Cantor model, `K80`* and `K81`* for the Kimura 2-parameters and Kimura 3-parameters models, and use no $(*)$ symbol for their original continuous-time versions.

Definition 1.15. The strand symmetric model, `SSM`, was first introduced in *Chapter 16* of Pachter and Sturmfels (2005b). It reflects the double strandedness of the DNA sequences. In the light of the findings of Yap and Pachter (2004), this model is proposed as its transition probabilities support complementary base pairing– in the double helix of DNA hydrogen bonds are created between A and T, and C and G. Thus, the model assumes that the entries of a substitution matrix A satisfy: $A_{\mathtt{AA}} = A_{\mathtt{TT}}$, $A_{\mathtt{AC}} = A_{\mathtt{TG}}$, $A_{\mathtt{AG}} = A_{\mathtt{TC}}$, $A_{\mathtt{AT}} = A_{\mathtt{TA}}$, $A_{\mathtt{CA}} = A_{\mathtt{GT}}$, $A_{\mathtt{CC}} = A_{\mathtt{GG}}$, $A_{\mathtt{AG}} = A_{\mathtt{TC}}$ and $A_{\mathtt{CT}} = A_{\mathtt{GA}}$. The root distribution probability is also strand symmetric: $\pi_{\mathtt{A}} = \pi_{\mathtt{T}}$ and $\pi_{\mathtt{C}} = \pi_{\mathtt{G}}$, and the substitution matrices are given by

$$\begin{pmatrix} a & b & c & d \\ e & f & g & h \\ h & g & f & e \\ d & c & b & a \end{pmatrix}, \quad \text{with} \quad \begin{array}{c} a+b+c+d=1, \\ e+f+g+h=1. \end{array}$$

It is the model that is best suited to describe large data sets– provided that we dispose sufficient amount of data, base composition will reflect the rules of base-pairing. The Strand Symmetric model (SSM, Casanellas and Sullivant (2005)) can be considered a discrete-time version of the HKY model (Hasegawa et al., 1985) with equal distribution of the pairs of bases (A, T) and (C, G) at each node of the tree.

We refer to Greuel et al. (2003), Allman and Rhodes (2004b), and (Semple and Steel, 2003, chapter 8) for further background and references on the discrete-time models.

Definition 1.16. If a DNA sequence has evolved from another according to a substitution matrix A^e, then the number of substitutions per site that have occurred can be approximated by

$$l(e) = -\frac{1}{4}\log\det(A^e) \tag{1.2}$$

(see Barry and Hartigan (1987)). This is usually known as the *branch length* of edge e.

The unit of measure for branch length in the work contained in this thesis will be the expected number of substitutions per site. That is, if e is a directed edge of a branch length equal to 0.5, then 50% of the sites have undergone a substitution along the evolutionary process on edge e. In the case of stationary continuous-time models, the expected numer of substitutions per site coincides with $-tr(D_\pi Q\lambda_e t_e)$ if $A^e = \exp(Q\cdot\lambda_e t_e)$ and D_π is a diagonal matrix with entries corresponding to the stationary distribution π. If the stationary distribution is uniform (as is for JC69*, K80* and K81* models), this can be rewritten as $-\frac{1}{4}\log\det(A^e)$. Note that in all continuous-time models the branch length can be computed from the matrix $\exp(D_\pi Q^e)$, which has the same shape as the transition matrix A^e.

On the contrary to the previous models, the distribution of the bases of the SSM and GMM models varies among the nodes of $\mathcal{T}$, thus it is not stationary. Stationary distribution is oftentimes referred to as the *stable base distribution* as it imposes the assumption of compositional homogeneity between the nucleotide bases. Assumption of this form was shown to be restrictive and mislead the phylogenetic reconstruction (see Jermiin et al. (2004) and the references within). Allman and Rhodes (2006b) translated these concepts into the algebraic language and introduced the Algebraic Time Reversible (ATR) and Stable Base Distribution (SBD) models (see below). Reversibility implies that the probability of a substitution between ancestor and descendant nodes along $\mathcal{T}$ is independent of the direction of time. This in turn implies that the frequency of the bases is constant at all points in the divergence time. The definitions of these models are given by Allman and Rhodes (2006b) are:

Definition 1.17. The $\pi-$*Stable Base Distribution* (SBD) model is the most general stationary model with equilibrium vector π. That is, if π is a vector whose elements sum to 1, $\pi = (\pi_{\mathtt{A}}, \pi_{\mathtt{C}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}})$, we require that $A^e\mathbf{1} = \mathbf{1}$ and $\pi = \pi A^e \quad \forall \quad e \in E(\mathcal{T})$.

Definition 1.18. Given a distribution π, the $\pi-$*Algebraic Time Reversible* (ATR) model is defined by the following conditions

(1) $A^e\mathbf{1} = \mathbf{1} \quad \forall e \in E(\mathcal{T})$,

(2) the set of matrices $(A^e)_{e\in E(\mathcal{T})}$ commute with each other, and

(3) the matrices $(D_\pi A^e)_{e\in E(\mathcal{T})}$ are symmetric.

The JC69*,K80*, K81* models are examples of the ATR model and thus of SBD. Among the models described above, we can write down the following chain of inclusions:

$$\begin{aligned} &\mathtt{JC69}^* \subset \mathtt{K80}^* \subset \mathtt{K81}^* \subset \mathtt{SSM} \subset \mathtt{GMM}, \\ &\mathtt{JC69}^* \subset \mathtt{K80}^* \subset \mathtt{K81}^* \subset \mathtt{ATR} \subset \mathtt{SBD} \subset \mathtt{GMM}. \end{aligned} \tag{1.3}$$

The lemma below states that the K81* matrices (as well as JC69*, K80*) are diagonalizable.

Lemma 1.19. *Let* $A = \begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix}$ *be a* K81* *matrix* ($a+b+c+d=1$) *and consider the matrix*

$$S = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \end{pmatrix}.$$

Then $S^{-1} = \frac{1}{4}S$ *and* $S^{-1}AS$ *is a diagonal matrix with diagonal entries* $\{1, a-b-c+d, a-b+c-d, a+b-c-d\}$ *(in this order).*

Remark 1.20. The change of variables considered in the Proposition above corresponds to the discrete Fourier transform in the setting of Sturmfels and Sullivant (2005).

The transition matrices of the JC69*, K80*, K81* and the SSM models show certain symmetries in their structures. This property allows to redefine them as equivariant models (see Section 5.5). The two latter models, SBD and ATR, give explicit conditions to test the important conditions of the evolutionary process, the reversibility and stationarity.

We will explore the properties of these models and the applications of new approaches to their analysis in modern phylogenetics throughout the rest of this work.

1.4.3 Invariants

The term *phylogenetic invariants* was coined Cavender and Felsenstein (1987) and Lake (1987) to name polynomial equations that could be used for phylogenetic reconstruction of tree topology. The definition evolved and now we distinguish different types of invariants.

For every tree $\mathcal{T}$, given a model and assigned transition probabilties to the edges, we can write the formula for the joint probabilities (see (1.1)). Given a tree $\mathcal{T}$ and a model $\mathcal{M}$, an *invariant* is a polynomial that vanishes on the expected joint distribution at the leaves of $\mathcal{T}$, $p^{\mathcal{T}}_{\mathcal{M},\mathtt{x}_1\ldots\mathtt{x}_n}(\pi,(A^e)_{e\in E(\mathcal{T})})$, irrespective of the choice of the continuous parameters (the substitution matrices and the root distribution).

Phylogenetic invariants are the invariants that can distinguish between different tree topologies: they vanish on all the joint probabilities for a given tree topology, but not on all for another tree. The work on phylogenetic invariants has been pursued by Allman and Rhodes (2006a), Allman and Rhodes (2008b), Draisma and Kuttler (2009), Casanellas and Sullivant (2005), Casanellas and Fernández-Sánchez (2011), to name a few.

It was already noted by e.g. Eriksson (2005), Sturmfels and Sullivant (2005) that some invariants depend only on the model chosen (and not on the topology). For example, in the case of the JC69* on a 3-star tree these include $\sum p_{\mathtt{x}_1\mathtt{x}_2x_3}=1$, $p_{\mathtt{AAA}}-p_{\mathtt{CCC}}=0$, $p_{\mathtt{CCC}}-p_{\mathtt{GGG}}=0$,$p_{\mathtt{GGG}}-p_{\mathtt{TTT}}=0$ (see Chap. 6.3 for the complete list). The polynomials that vanish on all $p^{\mathcal{T}}_{\mathcal{M},\mathtt{x}_1\ldots\mathtt{x}_n}$ for a given model $\mathcal{M}$ irrespective of the underlying tree topology are called *model invariants*.

The joint pattern frequencies can be estimated from the observed alignment. Given a multiple sequence alignment of n species, we can estimate the probability of occurrence of pattern $\mathtt{x}_1\mathtt{x}_2\ldots\mathtt{x}_n$ by the relative frequency of column $\mathtt{x}_1\mathtt{x}_2\ldots\mathtt{x}_n$ in the alignment. Thus, the relative frequencies of the columns of the multiple sequence alignments become a plug-in estimate for the polynomial equations that characterize a model and a tree topology, i.e. if the data evolved under a model $\mathcal{M}$, the model invariants for $\mathcal{M}$ are close to zero on these observed frequencies. We write "close to zero", as the data is limited and the theoretical "vanishing" of the invariants will not be attained in practical applications.

First we note that an undisputable advantage of the approach based on invariants is that it is parameter-free, i.e. the topology or the model are chosen without the need to estimate the superfluous parameters– invariants contain no obsolete information. Also, they are applicable to the nonhomogeneous models. That said, this new approach is far from perfect and its applicability is limited by a series of problems. Firstly, one needs an efficient way of listing the invariants. In some instances, for the models of low dimension and small trees, these invariants can be obtained of such computer algebra systems include: Singular (Greuel et al., 2001), CoCoa (Abbott et al., 2007; Abbott and Bigatti, 2010), Macaulay 2 (Grayson and Stillman, 2009) are examples of these. For large problems (e.g. large trees) the cardinality of the "sufficient" set of invariants will grow exponentially in n. Therefore a generating set of the "most relevant" (topologically or model informative) invariants should be obtained. The invariants were shown to outperform some alternative methods in tree reconstruction for 4−taxon trees (Casanellas and Fernández-Sánchez, 2007; Casanellas and Fernández-Sánchez, 2011), but for larger problems the are too computationally expensive. However, as shown in section 9, they offer an appealing framework for model selection in phylogenetic mixtures.

Chapter 2

Case-study on the conservation of splicing regulators

Illustrated in the context of splicing, this section describes the main motivations which led to the development of great part of the work presented in the thesis.

The content of the section is self-contained and not necessary for understanding concepts and results presented in the remaining chapters. However, we believe it provides valuable insights into usefulness and possible applications of the methodologies introduced in the following chapters.

It is being increasingly appreciated that the genomic sequence is intrinsically polysemic: the same DNA sequence often carries multiple meanings, i.e. it is involved in different functions. The nucleotide sequence of the genome, therefore, is shaped by multiple contrasting evolutionary forces acting at different levels (see e.g. Hurst (2006), Warnecke and Hurst (2007); Warnecke et al. (2008a,b, 2009),Washietl et al. (2008), Tilgner et al. (2009), Tilgner and Guigó (2010), Fairbrother et al. (2004), Carlini and Genut (2006)).

We show that diferent parts of coding sequences are subject to different selective constraints. This justifies the use of *mixtures of trees* in phylogenetic inference (see Def. 6.1). Within protein coding regions, sequences may play a role in control of translation, translational efficiency, transcript stability, etc. (see Chamary et al. (2006) for a review) and may therefore be subjected to additional selective forces not directly related to protein coding function. Sequences involved in the definition of splice sites, and in the regulation of alternative splicing (the ESEs, ESSs, ISEs and ISSs; see section 1.1) are examples of such. There is a stronger evidence that splicing regulatory sequences are under additional selective pressure in coding regions (Parmley et al. (2006), Orban and Olah (2001)). Neutrally evolving sequences are widely used to estimate divergence times between species. Oftentimes, the four fold degenerate positions within coding exons are used for this purpose. However, the assumption of evolutionary neutrality on these positions has been challenged upon realizing that many exonic sequences play functional roles not directly related to protein coding function. By analyzing human mouse orthologous gene pairs Parmley et al. (2006) show the rate of synonymous substitutions is lower in putative ESE sequences than in non-ESE sequences (see Remark 1.4).

We investigated whether extending this analysis by considering simultaneously orthologous constitutive exons across six different vertebrates would confirm the results and possibly contribute to gaining additional information. Consistent with the results of Parmley et al. (2006) we show in the synonymous positions overlapping a core set of ESE sequences are more constrained that synonymous positions not overlapping them. We used multiple nucleotide sequence alignments of coding exon sequences across six vertebrate species to infer the rate of evolutionary change at base pair resolution. We specifically compared the rate of evolution at synonymous positions covered by known splice regulators and at synonymous positions not covered by them.

2.1 Data

All data sets used in this section are available at `http://genome.imim.es/datasets/ESEselection2008/`.

Putative splicing regulators (ESE, ESS) Up to date the number of identified splicing-related regulatory subsequences comprises 78% of the total set of possible hexamers (6-tuples on the set $\{\texttt{A},\texttt{C},\texttt{G},\texttt{T}\}$). Thus, given their ubiquitousness defining a pertinent set of motifs acting in splicing is a nontrivial task. We used the list of 666 experimentally and computationally validated ESEs of Fairbrother et al. (2002) as a starting set. Next, we pruned them to derive a smaller set of 32 “trusted” regulatory pentamers (*ese*) by removing the first or last redundant base whenever a given hexamer had a wobble nucleotide in its first or last position, i.e. `AAAAA` was considered if either $\forall \texttt{x} \in \{\texttt{A},\texttt{C},\texttt{G},\texttt{T}\}$ `xAAAAA` or `AAAAAx` belonged to the original set of 666 hexamer ESEs (Tilgner and Guigó, 2007). As a“neutral” set we used a set of 886 hexamers that to our knowledge have not yet been implicated in splicing regulation. 60% of them was used in neutral model definition. The remaining 355 hexamers were used as a control test set (*nonESE*).

Multiple sequence alignments We chose five mammalian species: two primate species (human and macaque), two rodent species (mouse and rat), and an artyodactil (cow); and chicken as a relative outgroup. Figure 2.1 displays the placement of these species in a generally accepted species tree in the *ENCODE* project (see Section 1.2, Nikolaev et al. (2007)).

In order to obtain multiple alignments of orthologous exon sequences, we projected all *ENSEMBL* human transcripts (Hubbard et al. (2009)) onto the human genomic sequence and selected only coding internal exons longer than 146bp surrounded by fully intronic regions (non-terminal exons with consensus splice sites).

For these, we extracted the 70bp downstream of the acceptor $3'$ splice site, and the 70bp upstream of the $5'$ splice site skipping the 3 most proximal nucleotides to the splice sites.

We next identified the orthologous exons in the other species investigated. We used the *LiftOver* tool from the *USCS* Genome Browser (Fujita et al., 2010) to get the genomic positions corresponding to the human exons in *Rhesus macaque* (RheMac2), *Mus musculus* (Mm8), *Rattus norvegicus* (Rn4), *Bos Taurus* (BosTau3) and *Gallus*

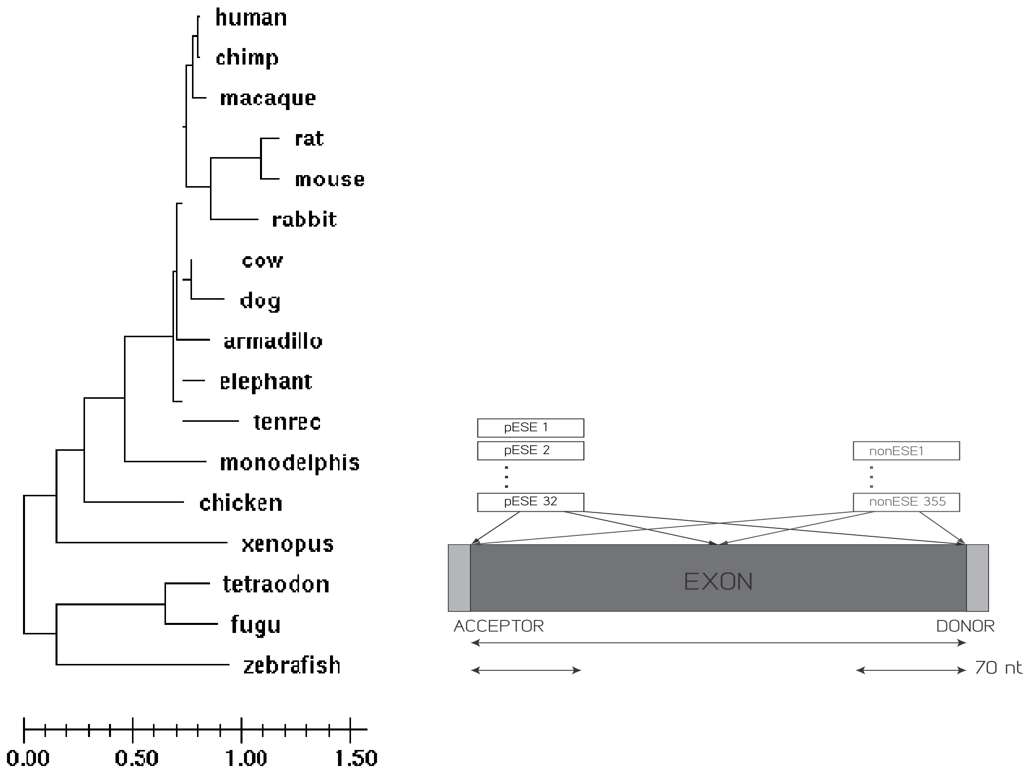


Figure 2.1: *Left*: Vertebrate tree derived from the four fold degnerate sites in genes from the ENCODE regions (Thomas et al. (2007)) using MAVID (Nikolaev et al. (2007)); *right*: placing regulatory and neutral motifs in the vicinity of the splice sites.

gallus (GalGal3). Only those exons with canonical splice sites (GT/AG) in all species were considered. In the end, we had a set of 8, 775 human constitutive exons conserved in all the species investigated.

For each of the 8, 775 sets of orthologous exons (orthologous exon groups), we performed an amino acid based nucleotide alignment. First, using human as reference, we inferred the phase of each of the orthologous exons in the other species. We translated each of the exons into all possible frames and kept the phase that gave the best score in a pairwise alignment with the human one using *T-coffee* (Notredame et al. (2000)). Then, we performed a multiple amino acid alignment of the exons, also using *T-coffee*, and translated it back to nucleotides. In this step, we removed all the orthologous exon groups containing "N's" or in frame stop codons. Finally, for each remaining orthologous exon group, we built exon-specific phylogenetic trees, and we retained for further analysis only those exon groups reproducing the established species phylogeny (as in Figure 2.1). We ended up with a set of 8, 583 alignments of "trusted" constitutive orthologous exons consisting in total of 1, 510, 077 alignment columns (orthologous coding nucleotide positions): 503, 370 corresponding to the first codon positions, 503, 350 to the second codon position and 503, 357 to the 3rd codon position. We extracted the subset of the 3rd codon positions that were synonymous across the entire alignment (227, 676 synonymous 3rd codon positions). The synonymous positions considered here were the four-fold degenerate sites, that is the third codon positions whose variations

Table 2.1: Number of alignment sites in the data sets used in the study.

defi						
defidata sets/		number of positions				
defi positions	number of exons	all	4−fold degenerate	(train)$nonESE$	(test)$nonESE$	ESE
defi full set	8,583	1,510,077	227,676	100,729	69,371	28,041
defi*weak 3' exons*	1,489	103,749	15,330	6,804	4,776	2,173
defi*strong 3' exons*	1,481	103,220	15,078	6,260	4,137	1,867
defi*weak 5' exons*	1,587	110,522	16,376	7,275	4,041	2,097
defi*strong 5' exons*	1,498	104,346	15,576	6,807	4,601	1,787

in the nucleotide does not affect the encoded amino acid.

In the end, we constructed a set with the four-fold degenerate sites from the 70 position next to the $3'$ (*3' exons*) and next to the $5'$ (*5' exons*). These sets are those in which all the analyses described in this paper was carried out. These includes the set of human constitutive exons, the orthologous exon groups, alignments and phylogenetic trees.

Exons We based our analysis on a set of constitutive coding human exons from the protein coding genes based on the evidence from the EST data. EST alignments were downloaded from *UCSC* (November 2007). We defined an exon as constitutive if it had at least a 90% inclusion level, i.e. at least 90% of the ESTs mapping to the exon region verify the exon. More formally, we define the inclusion level of an exon as $100\frac{N_i}{(N_i+N_e)}$, where N_i is the number of ESTs confirming the exon (EST verifies the exon boundaries $+/-$ 6nt), and N_e is the number of ESTs overlapping the region, but not including the exon. Only those exons exons with $N_i + N_e \geqslant 10$ were considered.

Splice site partitions of the orthologous exon alignments

We further divided the set of orthologous exon alignments according to the strength of their splice sites. We used standard Position Specific Scoring Matrices (PSSM) to score the splice sites. PSSMs for the acceptor ($3'$) and donor ($5'$) splice sites were derived from human splice sites. We pooled the scores of the splice sites from all species but chicken and identified the quartiles of the distribution. A splice site is defined to be “weak” if the score of the human splice site falls in the first quartile of the distribution whilst the corresponding scores in the remaining species do not exceed the second quartile. In a similar manner, a splice site is defined as “strong” if the score of the human splice site belongs to the fourth quartile (top 25% of the SS scores) and the scores for the remaining species lie in the second quartile (top 50%). The partition of exons was performed independently for $3'$ and $5'$ splice sites. This resulted in four subsets of the set of orthologous exon alignments: $1,489$ *weak 3' exons* ($103,749$ nt), $1,418$ *strong 3' exons* ($103,220$ nt), $1,587$ *weak 5' exons* ($110,522$ nt) and $1,498$ *strong 5' exons* ($104,346$ nt) (see Table 2.1 for a summary). Even though a given exon may belong to two different data sets (having a weak and a strong splice site), the nucleotide sequences extracted do not overlap.

Synonymous sites covered by regulatory pentamers and by neutral hexamers. We mapped the sets of *ESEs* , and neutral hexamers (the training and test sets) onto the human exon sequence (see Fig. 2.1). Only exact matches were considered. Next, we extracted the synonymous 3rd codon positions uniquely covered by any of the three sets: *ESEs* (28, 041 columns), by “training neutral” hexamers (100, 729 columns), and by the *nonESE* (693, 71 columns) (see Tab. 2.1).

Assessing sequence conservation. We measured conservation at each individual position based on a multiple nucleotide sequence alignment. Let $D = (D_1, \dots, D_N)$ be a multiple sequence alignment, where D_j denotes the jth column. We used a probabilistic measures of the conservation of D_j defined as:

$$\mathbf{p}(D_j) = -r_j \log{(r_j)}\,, \tag{2.1}$$

where $r_j = P(D_j \mid \mathcal{M}, \tau)$. Calculation of this score requires specification of an eovlutionary model $\mathcal{M}$ and a phylogenetic tree, $\mathcal{T}$. We made a heuristic choice as to the model and selected `HKY` model due to its flexibility. The parameters of the model $\mathcal{M}$ were estimated using *PAML* (Yang, 2007). Exemplary scores are given in Table 2.2. The score $\mathbf{p}$ takes 4^6 possible different real values. We then discretized the distribution (2.1) into m equally-spaced categories, such that m is the greatest integer smaller than $N^{\frac{1}{3}}$ (He and Meeden, 1997). Lastly, we use Kullback-Leibler divergence (KL) to quantify the distance between the distributions. Let $\Theta = (q_1, \dots, q_m)$ be the parameters associated to the positions evolving neutrally (*nonESE*) and $\Theta_{ese} = (p_1, \dots, p_m)$ of the positions covered by splicing regulators (*ese*). Both Θ and Θ_{ese} can be estimated as the relative frequencies of the score values falling within the m bins as observed in the data. The distance between the two distributions was measured by:

Table 2.2: Values of the phylogenetic conservation score with the parameters estimated for the set of positions in the vicinity of the $3'$ splice site.

human	A	A	A	A	A	A	C
macaque	A	A	A	A	A	C	A
rat	A	A	A	A	C	A	A
mouse	A	A	A	C	A	A	A
cow	A	A	C	A	A	A	A
chicken	A	C	A	A	A	A	A
Phylogenetic Conservation, *weak exons* [1]	0.2561	0.0551	0.0179	0.01	0.009	0.005	0.0037
Phylogenetic Conservation, *strong exons*	0.2866	0.0563	0.0175	0.0105	0.0098	0.0042	0.0035

$$-\operatorname{KL}(\Theta_{ese}, \Theta) = -\sum_{i=1}^{m} p_i \log \frac{p_i}{q_i}. \tag{2.2}$$

From the properties of KL it follows that the score takes values in $(-\infty, 0)$ and 0 indicating the equality of the two distributions. The exponent of this divergence belongs to $(0, 1)$ and can be interpreted as the probability that the set of alignment positions under consideration was generated under the “splicing-neutral” evolutionary model.

Henceforth, under this interpretation a value close to 1 indicates absence of negative selection (high divergence) and the lower the values the stronger the departure from neutrality.

2.2 Results and Discussion

Synonymous positions covered by Splicing Regulatory Sequences are more conserved than other synonymous position. As a reference set to both *ese* and *nonESE* sets we extracted the set of second codon positions from the multiple sequence alignments of exons under study.

As expected, second codon positions are more conserved than synonymous positions and the synonymous positions covered by ESE positions are more conserved than *nonESE* positions in both acceptor and donor data sets (see Fig. 2.2).

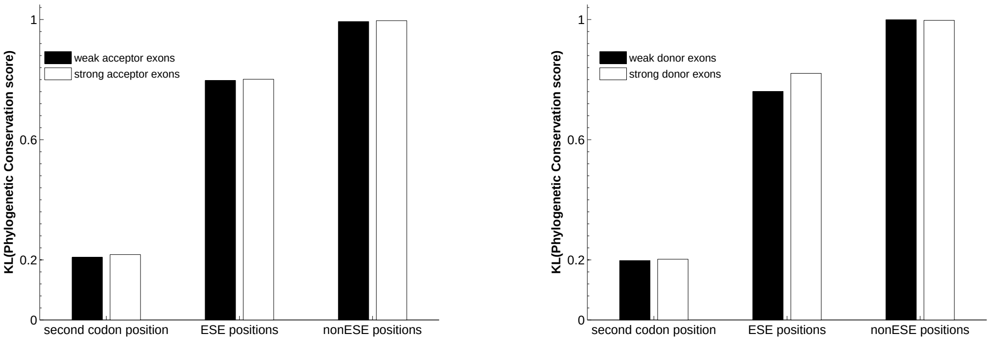


Figure 2.2: Exponent of the Kullback-Leibler divergence between the splicing-neutral training set of synonymous positions with *nonESEs* and the synonymos positions covered by *ese*, test set of *nonESE* and the second codon positions; see (2.2).

The average value of the score (2.2) taken across both splice sites was 0.2479 for the second codon position, 0.7659 for the synonymous ESE positions, *ese*, and 0.9956 for the synonymous *nonESE* positions.

In addition, as seen in Figure 2.2 there appears to be an small effect of the strength of donor splice sites, with ESE positions proximal to weak splice sites departing more from neutrality, unobserved in the vicinity of the acceptor sites.

The usage of synonymous positions covered by splicing regulatory sequences may confound estimates of evolutionary distance. Synonymous positions in coding regions are often used in tree inference. However, as shown in the results presented here are strongly indicative that (at least) a subset of synonymous positions in coding regions are not evolving neutrally. In particular, 4-fold positions within coding exons under selective constraints due to their role in the recognition of splice sites. It can be expected that using these constrained positions to estimate evolutionary distances will lead to an underestimation of divergence times.

We investigated the effect of ignoring the synonymous positions covered by the set *ese* in branch length estimation on the given *ENCODE* tree (see Fig.9.3). For this purpose we used *PAML* (Yang (2007)). First, we used all $229,796$ synonymous positions in the set of $8,583$ orthologoues coding exons to estimate the lengths of the branches.

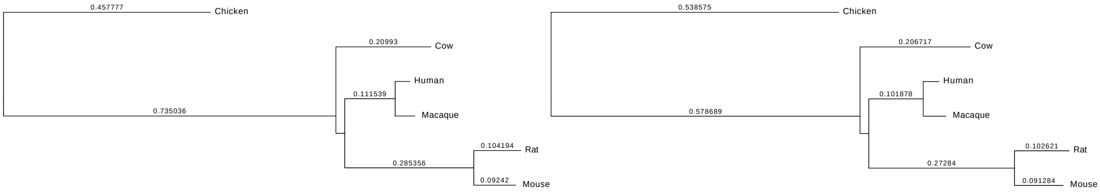


Figure 2.3: Branch length estimated from all 4-fold degenerate sites and from the positions with the synonymous sites covered by putative splicing regulators, *ese*, removed.

We then performed the same analysis by excluding the $35,783$ positions covered by *ese* ($194,013$nt) (see Fig. 9.3). As it is possible to see, with the exception of the branch leading to the chicken outgroup, all branches are slightly larger when the ESE positions are ignored -even though these constitute only a very small fraction of all synonymous positions. The total branch length of the tree computed as the sum of the branches of the phylogenetic tree is 1.98122 when using all synonymous positions and 2.09201 when excluding ESE positions.

In order to assess the statistical significance of the differences, we repeated the analysis 500 times. Each time we excluded the sets of synonymous positions covered random sets of 32 pentamers and calculated the tree length. As seen in Figure 2.4, the tree length obtained by excluding the positions covered by our set of confident ESEs (δ_{ese}) get noticeably longer than the trees obtained by excluding the positions covered by random sets of pentamers. The hypothesis is that the set of confident ESEs is indeed involved in regulation of splicing. In addition, there is additional selective pressures, not directly related too protein coding capacity, acting on the synonymous sites.

As other recent analysis (Ke et al. (2008), Goren et al. (2006)), our analysis shows that additional selective constraints are acting on protein coding regions not directly related to protein coding functionality. We have detected that selection is acting more strongly in synonymous third codon positions occurring in sequences that have been implicated in promoting exon inclusion (*ESEs*) than in those positions not occurring in such sequences (nonESE). We have been able to detect that selection may be slightly stronger in third codon positions proximal to weak donor sites than proximal to strong donor sites. In agreement with the findings of Xiao et al. (2007), no differences were observed between weak and strong acceptor sites, implying that the strength of the donor site is more important than that of the acceptor site to define splicing.

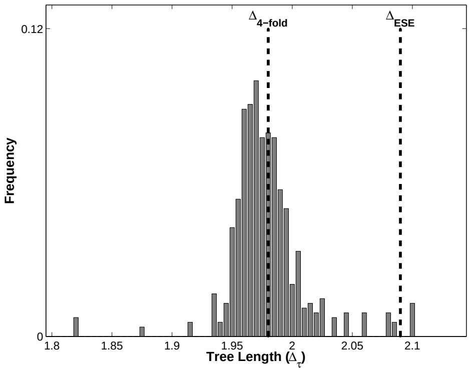


Figure 2.4: Length distribution of Phylogenetic trees relating the species investigated here (human, macaque, rat, mouse, cow and chicken) inferred from third codon synonymous positions after removing positions covered by random sets of 32 pentamers. The randomization was carried out 500 times. The lengths of the tree obtained from all four fold degenerate positions (Δ_{4fold}) and after removing the positions covered by the 32 ESE pentamers (ese: Δ_{ESE}) are depicted as vertical bar.

Enhancement in the detection of weak selective constraints can be attained by using the phylogenetic information relating the species—— when composite regions are analysed, mixture models are the most optimal choice. In addition, non-heuristic model selection motivated by the data at hand is an important pre-inference step in phylogenetics. Method for parameter estimate for complex (nonhomogeneous) data is also a question not fully addressed in the field. Motivated by the results presented in this section, in subsequent chapters we propose a framework for dealing with nonhomogenous models and their mixtures. Firstly, the model selected should not depend on the underlying tree. In fact, the data comprising concatenated set of divergent regions can be viewed as phylogenetic mixtures. In addition, an important step is surpassing the assumption of model homogeneity, i.e. allowing different rate matrices at different branches of the tree. One of the unanswered questions in phylogenomics is the number of divergent regions that could be concatenated for viable estimation. This question is related to statistical identifiability in maximum likelihood inference. Lastly, methods for branch length estimation in nonhomogeneous models are not yet established. This and other challenges posed here will be addressed in the progression of the work.

Part II

Algebraic tools in phylogenetics

Chapter 3

Markov evolutionary matrices for given branch lengths

Generating the DNA sequences evolving under a stationary continuous-time evolutionary model on an edge e with preassigned branch length l and given rate matrix Q, is not difficult: according to equation (1.2) one just needs to take $\lambda_e t_e = -l/tr(D_\Pi Q)$ and follow the usual process to generate a Poisson distribution according to these parameters. There are several programs available for generating data under most-used continuous-time evolutionary models, for example `seq-gen` (Rambaut and Grassly, 1997) and `evolver` in *PAML* (Yang, 2007). An extra effort is needed if the amount of "substitution events" , branch length, is fixed. We found that the problem of generating data under the more general discrete-time models is equivalent to generating substitution matrices A^e (belonging to the evolutionary model) with a given determinant. For the `JC69`*, `K81`*, `K80`* and `SSM` models (Propositions 3.1, 3.4, 3.6 and 3.17) the results are bidirectional and we provide algorithms for generating *any* strictly stochastic matrix M with determinant equal to a given number $K \in (0,1)$, when M is either a `JC69`*, `K81`*, `K80`* or `SSM` matrix. For the most general model `GMM` we provide a way of generating strictly stochastic matrices with determinant equal to K, but we are not able to claim whether we produce all of them. We observe that we are able to produce matrices that are not the exponential of a real rate matrix (cf. Remark 3.10).

Here we address the problem of providing stochastic matrices of the above shapes with given determinant $K \in (0,1)$. From the formula (1.2) we see that this is equivalent to generating substitution matrices for a branch of a given length. For the continuous-time stationary reversible models this is an easy task because the expected number of substitutions per site can be written down in terms of the trace of the rate matrix.

The algorithms proposed in this paper have been implemented in C++ in order to generate multiple sequence alignments of DNA data evolving on any phylogenetic tree (see section 7). Earlier version of the algorithms ws used for testing, `SPIn`, model selection method for phylogenetic mixturs (see Chap. 9 and Kedzierska et al. (2012)). An example of an algorithm to generate data on quartet trees under nonhomogeneous continuous-time models was given by Jermiin et al. (2003). Here and in the subsequent sections, we solve the problem in general setting, any tree and discrete-time model.

3.1 Generating discrete-time matrices with a given determinant

Generating JC69* matrices with a given determinant

Proposition 3.1. *Let $K \in (0,1)$ and let*

$$A = \begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix}, \quad a + 3b = 1,$$

be a JC69* *matrix. Then A is a strictly stochastic matrix with determinant equal to K if and only if $a = \frac{1}{4}(1 + 3K^{1/3})$, $b = \frac{1-a}{3}$.*

Proof. Using Lemma 1.19 we have $\det A = (\frac{4a-1}{3})^3$. Therefore, A has determinant equal to K if and only if $a = \frac{1}{4}(1 + 3K^{1/3})$. Moreover, as $K \in (0,1)$, we obtain $1 > a > 0$ (and so $0 < b = \frac{1-a}{3} < 1$), and we are done. □

Therefore we have:

Algorithm 3.2. (Generation of JC69* matrices with given determinant.)
Input: K in $(0,1)$.
Output: A strictly stochastic JC69* matrix A with determinant K.

```
Step 1: Set $a = \frac{1}{4}(1 + 3K^{1/3})$, $b = \frac{1-a}{3}$.

Final: Return
```

$$A = \begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix}.$$

Generating K80* matrices with a given determinant

Remark 3.3. As a technical step previous to the generation of K80* matrices with given determinant, we consider the polynomial

$$p_K(x) = -2x^3 + x^2 + K, \quad K \in (0,1),$$

and we observe that it has exactly one real root s which lies in $(\sqrt{K}, 1)$. Indeed, the coefficients of $p_K(x)$ have one variation in sign and those of $p_K(-x)$ have no variation in sign. Therefore, applying Descartes' rule we obtain that $p_K(x)$ has exactly one positive root s and no negative roots. Moreover, as K is a constant in $(0,1)$, we have that $p_K(\sqrt{K}) = 2K(1 - \sqrt{K})$ is positive and $p_K(1) = K - 1$ is negative, implying that s lies in $(\sqrt{K}, 1)$.

Using the formula for the roots of a cubic polynomial we obtain

$$s = \frac{1}{6} + \frac{1}{6}\sqrt[3]{1 + 54K + 6\sqrt{3K + 81K^2}} + \frac{1}{6}\sqrt[3]{1 + 54K - 6\sqrt{3K + 81K^2}}.$$

As a byproduct, the polynomial $p_K(-x)$ has exactly one real root which coincides with $-s$.

Proposition 3.4. *Let $K \in (0,1)$ and let s be the unique real root of $p_K(x) = -2x^3 + x^2 + K$ (see Remark 3.3). Let*

$$A = \begin{pmatrix} a & b & c & b \\ b & a & b & c \\ c & b & a & b \\ b & c & b & a \end{pmatrix},$$

be a K80* *matrix ($a + 2b + c = 1$), and consider the change of variables $\alpha = 1 - 2(b + c)$, $\beta = 1 - 4b$. Then A is a strictly stochastic matrix with determinant equal to K if and only if $\sqrt{K} < |\alpha| < s$ and $\beta = K/\alpha^2$.*

Proof. First we note that the inverse change of variables is $b = \frac{1-\beta}{4}$, $c = \frac{1+\beta-2\alpha}{4}$. Moreover, $\alpha = a - c$ and $\beta = a - 2b + c$ are the diagonal entries in $S^{-1}AS$ (different than 1) in Lemma 1.19 and therefore $\det(A) = \alpha^2\beta$.

$\Rightarrow$) Assume that A is strictly stochastic with determinant K. Then b is strictly positive, so that $\beta < 1$. As $K = \det(A) = \alpha^2\beta$ and $\beta < 1$, we obtain $|\alpha| > \sqrt{K}$. In particular, $\alpha \neq 0$ and we can write $\beta = K/\alpha^2$.

Using the inverse change of variables above and $\beta = K/\alpha^2$ we have

$$a > 0 \Leftrightarrow 2b + c < 1 \Leftrightarrow \frac{3 - K/\alpha^2 - 2\alpha}{4} < 1 \Leftrightarrow p_K(-\alpha) > 0.$$

As noted in Remark 3.3, $p_K(-x)$ has exactly one negative root which equals $-s$ and lies in $(-1, -\sqrt{K})$. As $p_K(-x)$ has positive leading term, $p_K(-\alpha) > 0$ only holds if $\alpha > -s$.

Similarly, c is strictly positive if and only if $p_K(\alpha) > 0$. Following an analogous argument, we obtain that $p_K(\alpha) > 0$ if and only if $\alpha < s$. Putting all together we obtain $\sqrt{K} < |\alpha| < s$, as desired.

$\Leftarrow$) Assume that $\sqrt{K} < |\alpha| < s$ and $\beta = K/\alpha^2$. In particular, we have $< \beta < \frac{K}{K} = 1$ and we obtaing that $b = \frac{1-\beta}{4}$ is strictly positive.

Now, as in the proof of $\Rightarrow$) we have that $c > 0$ if and only if $p_K(\alpha) > 0$. And also as above, this happens if and only if $\alpha < s$. As we assumed $|\alpha| < s$, we obtain $c > 0$.

Lastly, $a > 0$ if and only of $p_K(-\alpha) > 0$, and this holds if and only if $\alpha > -s$ (see proof of $\Rightarrow$). As we assumed $|\alpha| < s$, we get that A is a strictly stochastic matrix.

Moreover, $\det(A) = \alpha^2\beta = K$ as wanted. □

Using the previous result, we provide the following algorithm for generating strictly stochastic K80* matrices with given determinant K. It is worth pointing out that with this algorithm we are generating *all* `K80`* strictly stochastic matrices with determinant K.

Algorithm 3.5. (Generation of `K80`* matrices with given determinant.)
Input: K in $(0,1)$.
Output: A strictly stochastic `K80`* matrix A with determinant K.

```
Step 1: Compute the unique real root $s$ of $p_K(x)$ using Remark 3.3.

Step 2: Choose $\alpha$ randomly such that $\sqrt{K} < |\alpha| < s$.

Step 3: Let $\beta := K/\alpha^2$, $b := \frac{1-\beta}{4}$, $c := \frac{1+\beta-2\alpha}{4}$, and $a := 1-2b-c$.

 Final: Return
```

$$A := \begin{pmatrix} a & b & c & b \\ b & a & b & c \\ c & b & a & b \\ b & c & b & a \end{pmatrix}.$$

Generating `K81`* matrices with a given determinant

Previously to dealing with the case of `K81`* matrices, for each real number K in $(0,1)$, we let s be the unique positive root of the polynomial

$$q_K(z) := z(z+1)^2 - 4K.$$

Indeed, according to Descartes' rules of signs, this polynomial has at most one positive root. Moreover, as $q_K(K) < 0$ and $q_K(1) > 0$, there is exactly one positive root s and it lies in $(K,1)$. Using the formula for the roots of a cubic polynomial we obtain

$$s = -\frac{2}{3} - \frac{1}{3}\sqrt[3]{-1-54K+6\sqrt{3K+81K^2}} - \frac{1}{3}\sqrt[3]{-1-54K-6\sqrt{3K+81K^2}}. \quad (3.1)$$

Proposition 3.6. *Let $K \in (0,1)$ and let s be the unique real root of $q_K(z) := z(z+1)^2 - 4K$. Let*

$$A = \begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix},$$

be a `K81`* matrix ($a+2b+c=1$), and consider the change of variables $\alpha = 1-2(b+c)$, $\beta = 1-2(b+d)$, $\gamma = 1-2(c+d)$. Then A is a strictly stochastic matrix with determinant equal to K if and only if $|\alpha| \in (s,1)$, $|\beta| \in \left(I_{|\alpha|}, J_{|\alpha|}\right)$ where

$$I_{|\alpha|} = \max\left\{ \frac{-1+|\alpha|+\sqrt{(1-|\alpha|)^2+\frac{4K}{|\alpha|}}}{2}, \frac{1+|\alpha|-\sqrt{(1+|\alpha|)^2-\frac{4K}{|\alpha|}}}{2} \right\},$$

$$J_{|\alpha|} = \min\left\{ \frac{1+|\alpha|+\sqrt{(1+|\alpha|)^2-\frac{4K}{|\alpha|}}}{2}, \frac{1-|\alpha|+\sqrt{(1-|\alpha|)^2+\frac{4K}{|\alpha|}}}{2} \right\},$$

and $\gamma = \frac{K}{\alpha\beta}$.

Remark 3.7. As the change of variables above is symmetric in b, c, d, the roles of these three variables can be exchanged in the previous Proposition.

Before proving this Proposition we need the following technical lemma.

Lemma 3.8. *Let K be a real number in $(0, 1)$, let s be the unique positive solution to $z(z+1)^2 - 4K = 0$, and consider the function*

$$f(x, y) = 1 - x - y + \frac{K}{xy}$$

defined over $\mathbb{R}^2 \setminus \{0\}$.Given $y > 0$, we consider the set

$$\Omega_y = \{x \in \mathbb{R} \mid x > 0, f(x, y) > 0, f(x, -y) > 0, f(-x, y) > 0, f(-x, -y) > 0\}.$$

Then Ω_y is not empty if and only if $y > s$. Moreover, if $x \in \Omega_y$ and $y < 1$, then x belongs to (I_y, J_y) where

$$I_y = \max\left\{ \frac{-1+y+\sqrt{(1-y)^2+\frac{4K}{y}}}{2}, \frac{1+y-\sqrt{(1-y)^2-\frac{4K}{y}}}{2} \right\} \quad \textit{and}$$

$$J_y = \min\left\{ \frac{1+y+\sqrt{(1-y)^2-\frac{4K}{y}}}{2}, \frac{1-y+\sqrt{(1-y)^2+\frac{4K}{y}}}{2} \right\}.$$

Proof. We fix $y > 0$, and we view f and g as functions on x. For $x > 0$ we can multiply f, g by x and define quadratic functions $\tilde{f}_y(x) := -x^2 + (1-y)x + K/y$ and $\tilde{g}_y(x) := x^2 + (1+y)x + K/y$ so that x belongs to Ω_y if and only if $x > 0$, $\tilde{f}_y(x) > 0$, $\tilde{f}_{-y}(x) > 0$, $\tilde{g}_y(x) > 0$ and $\tilde{g}_{-y}(x) > 0$.

Note that $\tilde{f}_y$ has discriminant $\Delta_1(y) = (1-y)^2 + \frac{4K}{y}$ and $\tilde{g}_y$ has discriminant $\Delta_2(y) = (1+y)^2 - \frac{4K}{y}$.

We observe that $\Delta_1(y) > 0$ for $y > 0$. Therefore $\tilde{f}_y(x) = 0$ has two real solutions $x_{1,L}(y) = \frac{1-y-\sqrt{\Delta_1(y)}}{2}$, $x_{1,R}(y) = \frac{1-y+\sqrt{\Delta_1(y)}}{2}$, and $\tilde{f}_y(x)$ is positive for x in $(x_{1,L}, x_{1,R})$. Note that $\sqrt{\Delta_1(y)} > |1-y|$ for $y > 0$, so $x_{1,L}(y)$ is negative and $x_{1,R}(y)$ is positive. Therefore, for $x > 0$ and $y > 0$, $\tilde{f}_y(x)$ is positive if and only if $x \in (0, x_{1,R}(y))$.

On the other hand, as $\tilde{f}_{-y}$ has negative leading coefficient, there exists x with $\tilde{f}_{-y}(x) > 0$ if and only if $\Delta_1(-y) > 0$. Note that $\Delta_1(-y)$ is positive for $y > 0$ if and only if $y > s$ (indeed, $\Delta_1(-y)$ coincides with $q_K(y)/y$).

Thus $\tilde{f}_{-y}(x) > 0$ has a solution for $x > 0$, if and only if $y > s$. Now for $x > 0, y > s$, the roots of $\tilde{f}_{-y}(x) = 0$ are $x_{1,L}(-y)$ and $x_{1,R}(-y)$. Clearly $x_{1,R}(-y)$ and $x_{1,L}(-y)$ are both positive for $y > s$. Therefore, for $x > 0$ and $y > 0$, we have $\tilde{f}_{-y}(x) > 0$ if and only if $y > s$ and $x \in (x_{1,L}(-y), x_{1,R}(-y))$.

Now we study the positivity of $\tilde{g}_y(x)$ for $x > 0$. Note that $\tilde{g}_y$ has discriminant $\Delta_1(-y)$. As the leading coefficient of $\tilde{g}_y$ is positive, we have that $\tilde{g}_y(x) > 0$ for all $y < s$ and $x \in \mathbb{R}$ (because in this case the discriminant is negative). Moreover, if $y > s$, the real roots of $\tilde{g}_y(x) = 0$ are $x_{2,L}(y) = \frac{-(1+y)-\sqrt{\Delta_1(-y)}}{2}$ and $x_{2,R}(y) = \frac{-(1+y)+\sqrt{\Delta_1(-y)}}{2}$. They are both negative so that $\tilde{g}_y(-x)$ is positive for all $y > s$ and $x > 0$.

We study the positivity of $\tilde{g}_{-y}(x)$ for $x > 0$ and $y > 0$. The discriminant of $\tilde{g}_{-y}$ is $\Delta_1(y)$, and it is positive for $y > 0$. Then the roots of $\tilde{g}_{-y}$ are $x_{2,L}(-y)$ and $x_{2,R}(-y)$. For $y > 0$ we have $x_{2,L}(-y) < 0$ and $x_{2,R}(-y) > 0$, and therefore $\tilde{g}_{-y}(x) > 0$ if and only if x belongs to $(x_{2,R}(-y), +\infty)$.

Summing up, we have proven that the set Ω_y is non-empty if and only if $y > s$. Moreover, in that case, if x belongs to Ω_y, then x lies in

$$(0, x_{1,R}(y)) \cap (x_{1,L}(-y), x_{1,R}(-y)) \cap (0, +\infty) \cap (x_{2,R}(-y), +\infty).$$

It is easy to see that $x_{1,R}(y)$ is bigger than $x_{2,R}(-y)$ for $y > 0$. Therefore the intersection of intervals above is equal to

$$(x_{1,L}(-y), x_{1,R}(-y)) \cap (x_{2,R}(-y), x_{1,R}(y)).$$

The statement of the lemma follows from the following claim.

Claim: If $y < 1$, then $x_{2,R}(-y) < x_{1,R}(-y)$.

Proof of Claim: This is equivalent to proving

$$\sqrt{\Delta_1(y)} - \sqrt{\Delta_1(-y)} < 2. \tag{3.2}$$

First of all we note that $\Delta_1(y) \leqslant \Delta_1(-y)$ if and only if $y \geqslant \frac{2K}{y}$. As $y > 0$, this holds if and only if $y \geqslant \sqrt{2K}$. Therefore, for $y \geqslant \sqrt{2K}$, $\sqrt{\Delta_1(y)} - \sqrt{\Delta_1(-y)}$ is negative (and hence < 2.)

If $y < \sqrt{2K}$, we have just seen that $\sqrt{\Delta_1(y)} > \sqrt{\Delta_1(-y)}$. In this case, both sides in (3.2) are positive and hence it is equivalent when raising it to the second power:

$$\Delta_1(y) + \Delta_1(-y) - 2\sqrt{\Delta_1(y)\Delta_1(-y)} < 4.$$

As we are assuming $y < 1$, we have $\Delta_1(y) + \Delta_1(-y) - 4 = 2y^2 - 2 < 0 < 2\sqrt{\Delta_1(y)\Delta_1(-y)}$, as we wanted to prove. □

Proof of Proposition 3.6. Taking into account that $a = 1-b-c-d$, we note that inverse change of variables is $a = \frac{1}{4}(1+\alpha+\beta+\gamma)$, $b = \frac{1}{4}(1-\alpha-\beta+\gamma)$, $c = \frac{1}{4}(1-\alpha+\beta-\gamma)$, $d = \frac{1}{4}(1+\alpha-\beta-\gamma)$. Observing that α, β, γ are the diagonal entries in $S^{-1}AS$ in Lemma 1.19, we see that $\det(A) = \alpha\beta\gamma$.

$\Rightarrow$) Assume that A is stochastic with determinant $K \in (0,1)$. Then α, β, and γ are non-zero, and $\gamma = \frac{K}{\alpha\beta}$. From the positivity of a, b, c, d we get that $1+\alpha+\beta+\frac{K}{\alpha\beta} > 0$, $1-\alpha-\beta+\frac{K}{\alpha\beta} > 0$, $1-\alpha+\beta-\frac{K}{\alpha\beta} > 0$, and $1+\alpha-\beta-\frac{K}{\alpha\beta} > 0$. In terms of Lemma

3.8, these inequalities can be rewritten as

$$f(-\beta,-\alpha)>0, f(\beta,\alpha)>0, f(\beta,-\alpha)>0, f(-\beta,\alpha)>0.$$

Therefore $|\beta|$ is an element of $\Omega_{|\alpha|}$, which implies that $|\alpha| > s$ (see Lemma 3.8). Moreover, as $\alpha = 1-2(b+d)$, and $b, d > 0$, we see that $|\alpha| < 1$. The result then follows from Lemma 3.8.

$\Leftarrow$) Using Lemma 3.8 we see that under these assumptions, $\Omega_{|\alpha|} \neq \emptyset$ and $|\beta|$ belongs to $\Omega_{|\alpha|}$. Therefore $f(-\beta,-\alpha)>0, f(\beta,\alpha)>0, f(\beta,-\alpha)>0, f(-\beta,\alpha)>0$. As $\gamma = \frac{K}{\alpha\beta}$, these inequalities coincide with $a > 0$, $b > 0$, $c > 0$ and $d > 0$, and we are done. □

The previous results give us a way of generating *any* K81* matrix.

Algorithm 3.9. (Generation of K81* matrices with given determinant.)
Input: K in $(0,1)$.
Output: A strictly stochastic K81* matrix A with determinant K.

```
Step 1: Compute the unique real root $s$ of $z(z+1)^2-4K$using (3.1).

Step 2: Choose $\alpha$ randomly such that $1 > |\alpha| > s$.

Step 3: Take $\beta$ randomly such that $|\beta|$ belongs to $(I_{|\alpha|}, J_{|\alpha|})$.

Step 4: Set $\gamma = \frac{K}{\alpha\beta}$.

Step 5: Set $a=\frac{1}{4}(1+\alpha+\beta+\gamma)$, $b=\frac{1}{4}(1-\alpha-\beta+\gamma)$, $c=\frac{1}{4}(1-\alpha+\beta-\gamma)$, $d=\frac{1}{4}(1+\alpha-\beta-\gamma)$.

 Final: Return
```

$$A = \begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix}.$$

Remark 3.10. The change of variables in Proposition 3.6 diagonalizes the matrix to $\mathrm{Diag}(1,\alpha,\beta,\gamma)$ (see Lemma 1.19). As we have seen in that proposition, α and β can be both negative. Therefore, using Culver (1966), we observe that the matrices produced by the algorithm above are not all of them of type $\exp(Q)$ for a real matrix Q.

Generating SSM matrices with a given determinant

Definition 3.11. Let A be a 4×4 real matrix. We call $F(A)$ the matrix obtained from A after performing the basis change $F(A) = S^{-1}AS$ where

$$S = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

When A is a SSM matrix, A can be viewed as an element in $Hom_G(\mathbb{C}^4, \mathbb{C}^4)$ where $G =< (\mathtt{AT})(\mathtt{CG}) >$ (see Casanellas and Fernández-Sánchez (2007)). The change of basis above decomposes $\mathbb{C}^4$ into its isotypic components via the natural linear representation $G \longrightarrow GL(\mathbb{C}^4)$. This change of basis is also known as the generalized Fourier transform (see Casanellas and Sullivant (2005)). We have the following fact:

Lemma 3.12. *A* 4×4 *matrix* $A = (a_{i,j})$ *is a* SSM *matrix if and only if* $F(A)$ *has the following shape:*

$$F(A) = \begin{pmatrix} \lambda & 1-\lambda & 0 & 0 \\ 1-\mu & \mu & 0 & 0 \\ 0 & 0 & \alpha & \alpha' \\ 0 & 0 & \beta' & \beta \end{pmatrix}.$$

In this case, $\lambda, \mu, \alpha, \alpha', \beta, \beta'$ *can be written in terms of the entries of* A *as* $\lambda = a_{1,1}+a_{1,4}$, $\mu = a_{2,2} + a_{2,3}$, $\alpha = a_{2,2} - a_{2,3}$, $\alpha' = a_{2,4} - a_{2,1}$, $\beta = a_{1,1} - a_{1,4}$, *and* $\beta' = a_{1,3} - a_{1,2}$. *The inverse change of variables is* $a_{1,1} = (\lambda + \beta)/2$, $a_{1,2} = (1 - \lambda - \beta')/2$, $a_{1,3} = (1-\lambda+\beta')/2$ $a_{1,4} = (\lambda-\beta)/2$, $a_{2,1} = (1-\mu-\alpha')/2$, $a_{2,2} = (\mu+\alpha)/2$, $a_{2,3} = (\mu-\alpha)/2$, $a_{2,4} = (1 - \mu + \alpha')/2$.

Proof. The matrix $F(A)$ for a generic matrix $A = (a_{i,j})$ is

$$\frac{1}{2}\left(\begin{array}{cc|cc} a_{1,1}+a_{1,4}+a_{4,1}+a_{4,4} & a_{1,2}+a_{1,3}+a_{4,2}+a_{4,3} & a_{1,2}-a_{1,3}+a_{4,2}-a_{4,3} & a_{1,4}-a_{1,1}-a_{4,1}+a_{4,4} \\ a_{2,1}+a_{2,4}+a_{3,1}+a_{3,4} & a_{2,2}+a_{2,3}+a_{3,2}+a_{3,3} & a_{2,2}-a_{2,3}+a_{3,2}-a_{3,3} & a_{2,4}-a_{2,1}-a_{3,1}+a_{3,4} \\ \hline a_{2,1}+a_{2,4}-a_{3,1}-a_{3,4} & a_{2,2}+a_{2,3}-a_{3,2}-a_{3,3} & a_{2,2}-a_{2,3}-a_{3,2}+a_{3,3} & a_{2,4}-a_{2,1}+a_{3,1}-a_{3,4} \\ a_{4,1}+a_{4,4}-a_{1,1}-a_{1,4} & a_{4,2}+a_{4,3}-a_{1,2}-a_{1,3} & a_{1,3}-a_{1,2}+a_{4,2}-a_{4,3} & a_{1,1}-a_{1,4}-a_{4,1}+a_{4,4} \end{array}\right).$$

If A is a SSM matrix, then $a_{3,1} = a_{2,4}$, $a_{3,2} = a_{2,3}$, $a_{3,3} = a_{2,2}$, $a_{3,4} = a_{2,1}$, $a_{4,1} = a_{1,4}$, $a_{4,2} = a_{1,3}$, $a_{4,3} = a_{1,2}$, and $a_{4,4} = a_{1,1}$. Therefore the non-diagonal blocks are 0. Moreover, as sums of rows are equal to 1, we have that the entries of each row in the upper left block sum to 1:

$$\frac{1}{2}(a_{1,1} + a_{1,4} + a_{4,1} + a_{4,4} + a_{1,2} + a_{1,3} + a_{4,2} + a_{4,3}) = 1,$$

$$\frac{1}{2}(a_{2,1} + a_{2,4} + a_{3,1} + a_{3,4} + a_{2,2} + a_{2,3} + a_{3,2} + a_{3,3}) = 1.$$

Conversely, imposing that the entries of non-diagonal blocks in $F(A)$ are equal to 0 is equivalent to imposing $a_{3,1} = a_{2,4}$, $a_{3,2} = a_{2,3}$, $a_{3,3} = a_{2,2}$, $a_{3,4} = a_{2,1}$, $a_{4,1} = a_{1,4}$, $a_{4,2} = a_{1,3}$, $a_{4,3} = a_{1,2}$, and $a_{4,4} = a_{1,1}$ (adding and subtracting certain pairs of equations). Moreover, $F(A)_{1,1} + F(A)_{1,2} = 1$ implies that sum of rows 1 and 4 is equal to 2 (and similar for rows 2 and 3). But we have just seen that the set of entries in the first (resp. second) row is equal to the set of entries in the forth (resp. third) row, thus the sum of entries in each row is equal to 1. □

In the following lemma we characterize the stochasticity of A via $F(A)$.

Lemma 3.13. *A is a strictly stochastic* SSM *matrix if and only if*

$$F(A) = \begin{pmatrix} \lambda & 1-\lambda & 0 & 0 \\ 1-\mu & \mu & 0 & 0 \\ 0 & 0 & \alpha & \alpha' \\ 0 & 0 & \beta' & \beta \end{pmatrix}$$

with $\lambda, \mu \in (0,1)$, $|\beta| < \lambda$, $|\beta'| < 1-\lambda$, $|\alpha| < \mu$, and $|\alpha'| < 1-\mu$.

Proof. If A is a SSM matrix, then

$$A = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ h & g & f & e \\ d & c & b & a \end{pmatrix}$$

with $a+b+c+d = 1$, $e+f+g+h = 1$, and by Lemma 3.12, $F(A)$ has the shape above with $\lambda = a+d$, $\mu = g+f$, $\beta = a-d$, $\beta' = c-b$, $\alpha = f-g$, and $\alpha' = h-e$.

If $a, b, \ldots, h$ are strictly positive, then we clearly have $\lambda, \mu \in (0,1)$, $|\alpha| < \mu$,$|\alpha'| < 1-\mu$, $|\beta| < \lambda$, and $|\beta'| < 1-\lambda$.

Conversely, if $F(A)$ is block-diagonal as in the statement of the lemma, we know by Lemma 3.12 that A is a SSM matrix with entries as above. As the inverse change of variables is $a = (\lambda+\beta)/2$, $b = (1-\lambda-\beta')/2$, $c = (1-\lambda+\beta')/2$ $d = (\lambda-\beta)/2$, $e = (1-\mu-\alpha')/2$, $f = (\mu+\alpha)/2$, $g = (\mu-\alpha)/2$, $h = (1-\mu+\alpha')/2$, then if λ, μ lie $(0,1)$, $|\alpha| < \mu$, $|\alpha'| < 1-\mu$, $|\beta| < \lambda$, and $|\beta'| < 1-\lambda$, we obtain that $a, b, \ldots, h$ are strictly positive. □

Before stating the main result of this section we introduce some notation and we prove a technical result.

Remark 3.14. Given $K \in (0,1)$, we consider the polynomial $r_K(z) = z^3 + z - 2K$. It has a unique positive real root. Indeed, by Descartes' rule of signs we see that r_K has at most one positive real root. Moreover, as $r_K(K)$ is strictly negative and $r_K(1)$ is strictly positive, there exists exactly one positive root ν_0 of $r_K(z)$ and it lies in $(K,1)$. Using the formula for the roots of a cubic polynomial we actually get

$$\nu_0 = -\frac{1}{3}\sqrt[3]{-27K + 3\sqrt{81K^2+3}} - \frac{1}{3}\sqrt[3]{-27K - 3\sqrt{81K^2+3}}.$$

Definition 3.15. Given $K \in (0,1)$, we consider the polynomial $r_K(z) = z^3 + z - 2K$ and we call ν_0 its unique positive root (Remark 3.14). We define Θ as the set of points $(\lambda, \mu) \in (0,1)^2$ satisfying

$$\nu_0 + 1 \leqslant \lambda + \mu < 2, \text{ and } |\lambda - \mu| < \min\left\{2 - \lambda - \mu, \sqrt{\frac{r_K(\lambda+\mu-1)}{\lambda+\mu-1}}\right\}.$$

Lemma 3.16. *Let λ, μ be real numbers in $(0,1)$ with $\lambda+\mu>1$. Then (λ,μ) belongs to Θ if and only if*

$$\frac{K}{\lambda+\mu-1}-(1-\lambda)(1-\mu)<\lambda\mu. \tag{3.3}$$

Proof. As $\lambda+\mu>1$, we exchange the inequality (3.3) by the following equivalent inequality:

$$(\lambda+\mu-1)(2\lambda\mu+1-\lambda-\mu)-K>0. \tag{3.4}$$

We consider the change of variables $s:=\lambda+\mu$, $t:=\lambda-\mu$ (so that $\lambda=\frac{s+t}{2}$, $\mu=\frac{s-t}{2}$). We observe that λ and μ lie in $(0,1)$ if and only if $|t|<s$ and $|t|<2-s$. As we are assuming $\lambda+\mu>1$, we have $s>2-s$. Therefore, λ,μ are real numbers in $(0,1)$ with $\lambda+\mu>1$ if and only if $|t|<2-s$.

In these new variables inequality (3.4) reads as $(s-1)(\frac{s^2-t^2}{2}+1-s)-K>0$, which is equivalent to

$$t^2<\frac{(s-1)((s-1)^2+1)-2K}{s-1}=\frac{r_K(s-1)}{s-1}. \tag{3.5}$$

$\Leftarrow$) Let λ,μ be real numbers in $(0,1)$ satisfying $\lambda+\mu>1$ and (3.4). Then $s:=\lambda+\mu$ lies in $(1,2)$, $|t:=\lambda-\mu|<2-s$, and s,t satisfy (3.5). In particular, $\frac{r_K(s-1)}{s-1}\geqslant 0$. As we have $s>1$, this inequality is positive if and only if its numerator is positive, which holds if and only if $s-1\geqslant\nu_0$. Therefore s is in $[\nu_0+1,2)$ and $|t|<\min\left\{2-s,\sqrt{\frac{r_K(s-1)}{s-1}}\right\}$; in other words, (λ,μ) belongs to Θ.

$\Rightarrow$) Conversely, let $(\lambda,\mu)\in\Theta$. Then, using the change of variables above, we have that (s,t) satisfies $|t|<\sqrt{\frac{r_K(s-1)}{s-1}}$. In particular, (3.5) is satisfied and hence (3.3) is satisfied as well. □

Proposition 3.17. *Given K a real number in $(0,1)$, we consider the polynomial $r_K(z)=z^3+z-2K$ and let ν_0 be its positive real root in $(K,1)$ (see Remark 3.14). We fix two real numbers λ,μ in $(0,1)$ such that $\lambda+\mu>1$. Then the set*

$$\Omega_{\lambda,\mu}=\left\{(\alpha,\beta)\in\mathbb{R}^2\,\middle|\,0<\alpha<\mu,|\beta|<\lambda,|\alpha\beta-\frac{K}{\lambda+\mu-1}|<(1-\lambda)(1-\mu)\right\}$$

is non-empty if and only if (λ,μ) belongs to Θ. Moreover in this case, (α,β) belongs to $\Omega_{\lambda,\mu}$ if and only if α belongs to $\left(\frac{\frac{K}{\lambda+\mu-1}-(1-\lambda)(1-\mu)}{\lambda},\mu\right)$, $\alpha>0$, and

$$\max\left\{-\lambda,\frac{\frac{K}{\lambda+\mu-1}-(1-\lambda)(1-\mu)}{\alpha}\right\}<\beta<\min\left\{\lambda,\frac{\frac{K}{\lambda+\mu-1}+(1-\lambda)(1-\mu)}{\alpha}\right\}.$$

Proof. $\Rightarrow$) If (α,β) is a point in $\Omega_{\lambda,\mu}$, then $|\alpha\beta-\frac{K}{\lambda+\mu-1}|<(1-\lambda)(1-\mu)$. This is equivalent to

$$\frac{K}{\lambda+\mu-1}-(1-\lambda)(1-\mu)<\alpha\beta<\frac{K}{\lambda+\mu-1}+(1-\lambda)(1-\mu). \tag{3.6}$$

In particular, as $\alpha\beta < \lambda\mu$, we have

$$\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu) < \lambda\mu.$$

Hence, using Lemma 3.16 we obtain $(\lambda,\mu) \in \Theta$.

Moreover, as $|\beta| < \lambda$, inequality $\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu) < \alpha\beta$ implies $\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu) < \lambda\alpha$, and therefore α belongs to the interval

$$\left(\frac{\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu)}{\lambda}, \mu\right).$$

The inequalities on β follow directly from (3.6) and from $|\beta| < \lambda$. Conversely, if α belongs to the above interval, and β satisfies

$$\max\left\{-\lambda, \frac{\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu)}{\alpha}\right\} < \beta < \min\left\{\lambda, \frac{\frac{K}{\lambda+\mu-1} + (1-\lambda)(1-\mu)}{\alpha}\right\},$$

then inequalities (3.6) hold and hence (α,β) lies in $\Omega_{\lambda,\mu}$.

$\Leftarrow$) Let (λ,μ) be a point in Θ. In this case (λ,μ) satisfies (3.3), and in particular, the interval

$$\left(\frac{\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu)}{\lambda}, \mu\right) \tag{3.7}$$

is non-empty. We choose $\alpha > 0$ in this interval.

Then, the interval

$$\left(\frac{\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu)}{\alpha}, \frac{\frac{K}{\lambda+\mu-1} + (1-\lambda)(1-\mu)}{\alpha}\right)$$

is non-empty (the left-hand side numerator is smaller than the right-hand side numerator, and the denominator is positive) and its intersection with $(-\lambda,\lambda)$ is not empty. Indeed, as $\alpha > 0$ and α belongs to the interval (3.7), we have

$$\frac{\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu)}{\alpha} < \lambda;$$

moreover $-\lambda$ is less than $\frac{\frac{K}{\lambda+\mu-1}+(1-\lambda)(1-\mu)}{\alpha}$ because this expression is positive.

Finally, we choose β in this intersection of intervals and we obtain a point (α,β) in $\Omega_{\lambda,\mu}$. □

Theorem 3.18. *Let K be a real number in $(0,1)$.*

(a) Let (λ,μ) be a point in Θ, let (α,β) be a point in $\Omega_{\lambda,\mu}$, and consider real numbers α' and β' such that

(i) $\frac{|\alpha\beta - \frac{K}{\lambda+\mu-1}|}{1-\mu} < |\beta'| < 1-\lambda$, and

(ii) $\alpha' = \frac{\alpha\beta - \frac{K}{\lambda+\mu-1}}{\beta'}$.

Then, if we consider the change of variables $a = (\lambda+\beta)/2$,$b = (1-\lambda-\beta')/2$,$c = (1-\lambda+\beta')/2$ $d = (\lambda-\beta)/2$, $e = (1-\mu-\alpha')/2$, $f = (\mu+\alpha)/2$, $g = (\mu-\alpha)/2$, $h = (1-\mu+\alpha')/2$, *the matrix*

$$A = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ h & g & f & e \\ d & c & b & a \end{pmatrix}$$

is a strictly stochastic SSM *matrix with determinant* K, $a+d+f+g > 1$, $b \neq c$, *and* $f < g$.

(b) *Conversely, let*

$$A = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ h & g & f & e \\ d & c & b & a \end{pmatrix}$$

be a strictly stochastic SSM *matrix with determinant* K *and with* $a+d+g+f > 1$, $b \neq c$ *and* $f > g$. *Then* $F(A)$ *is equal to*

$$\begin{pmatrix} \lambda & 1-\lambda & 0 & 0 \\ 1-\mu & \mu & 0 & 0 \\ 0 & 0 & \alpha & \alpha' \\ 0 & 0 & \beta' & \beta \end{pmatrix},$$

where $(\lambda, \mu) \in \Theta$, $(\alpha, \beta) \in \Omega_{\lambda,\mu}$, *and* α', β' *satisfy conditions (i) and (ii) stated in (a).*

Remark 3.19. (1) By Proposition 3.17, if (λ, μ) is a point in Θ, there exists $(\alpha, \beta) \in \Omega_{\lambda,\mu}$. This implies that $|\alpha\beta - \frac{K}{\lambda+\mu-1}|$ is smaller than $(1-\lambda)(1-\mu)$, and thus the interval

$$\left(\frac{|\alpha\beta - \frac{K}{\lambda+\mu-1}|}{1-\mu}, 1-\lambda \right)$$

is non-empty. In particular, there exists β' in this interval. Therefore conditions (i) and (ii) in Theorem 3.18(a) are not empty.

(2) Assumptions $a+d+g+f > 1$, $f > g$, $b \neq c$ are biologically meaningful: the elements in the diagonal of an evolutionary Markov matrix stand for the conditional probabilities of no mutation, which are supposed to be much higher than the off-diagonal probabilities. It is even reasonable to assume that these diagonal entries are greater than 0.5, giving in particular $a+d+g+f > 1$. In any case, the result proved above can be easily adapted to the case $a+d+g+f < 1$ or $f > g$ (we have not done it here in order to make the paper more readable). Note also that any SSM matrix with determinant K and $f > g$ gives rise to a SSM matrix with $f < g$ and determinant K by permuting its 1st and 4th rows and its 2nd and 3rd rows (or columns, if preferred).

The hypothesis $b \neq c$ was added to simplify the statement of the Theorem and can be easily removed. Indeed, a matrix A as in (b) has $b = c$ and determinant equal to K if and only if $F(A)$ has $\beta' = 0$ and K is equal to $(\lambda + \mu - 1)\alpha\beta$. Therefore A is strictly stochastic with determinant K and $b = c$ if and only if $\frac{K}{\lambda(\lambda+\mu-1)} < |\alpha| < \mu$, $\beta = \frac{K}{\alpha(\lambda+\mu-1)}$, $\beta' =$ and α' is any number satisfying $|\alpha'| < 1 - \mu$.

Proof. (a) Let A be defined from $\lambda, \mu, \beta, \ldots, \alpha$ as above. Then $F(A)$ is equal to

$$B = \begin{pmatrix} \lambda & 1-\lambda & 0 & 0 \\ 1-\mu & \mu & 0 & 0 \\ 0 & 0 & \alpha & \alpha' \\ 0 & 0 & \beta' & \beta \end{pmatrix}.$$

We prove that A is a stochastic matrix using Lemma 3.13.

By hypothesis, $(\lambda, \mu) \in \Theta$ and hence λ and μ lie in $(0, 1)$. Moreover, as $(\alpha, \beta) \in \Omega_{\lambda,\mu}$, we have $0 < \alpha < \mu$, $|\beta| < \lambda$. By assumption (i), $|\beta'| < 1 - \lambda$ is also satisfied. It remains to prove that $|\alpha'| < 1 - \mu$. But this follows from conditions (i) and (ii):

$$|\alpha'| = \frac{|\alpha\beta - \frac{K}{\lambda+\mu-1}|}{|\beta'|} < 1 - \mu.$$

Row sums in A are equal to 1 by definition of $a, b, \ldots, h$. Moreover, as $B = F(A)$ is obtained from A by a basis change, we have that $\det A = \det B$ and it coincides with $(\lambda + \mu - 1)(\alpha\beta - \alpha'\beta')$. Thus, by assumption (ii) we have $\det A = K$.

(b) Lemma 3.12 tells us that $F(A)$ has the shape in the statement of the Proposition, and that $\lambda = a + d$, $\mu = g + f$, $\alpha = f - g$, $\alpha' = h - e$, $\beta = a - d$, and $\beta' = c - b$. By Lemma 3.13 we have that λ, μ lie in $(0, 1)$, $[\alpha| < \lambda$, $|\beta| < \lambda$, $|\alpha'| < 1 - \mu$, $|\beta'| < 1 - \lambda$. Moreover, as we are assuming $a + d + g + f > 1$, $b \neq c$, and $f > g$, we have $\lambda + \mu > 1$, $\beta' \neq 0$, and $0 < \alpha < \mu$.

On the other hand, $\det A = K$ implies $K = (\lambda + \mu - 1)(\alpha\beta - \alpha'\beta)$ and therefore condition (ii) holds.

The remaining inequality in (i),

$$\frac{|\alpha\beta - \frac{K}{\lambda+\mu-1}|}{1 - \mu} < |\beta'|,$$

holds because $|\alpha'|$ satisfies (ii) and $|\alpha'| < 1 - \mu$.

We prove now that (α, β) belongs to $\Omega_{\lambda\mu}$, that is,

$$|\alpha\beta - \frac{K}{\lambda + \mu - 1}| < (1 - \lambda)(1 - \mu). \tag{3.8}$$

We have just seen that $|\beta'|$ satisfies condition (i), so

$$|\alpha\beta - \frac{K}{\lambda + \mu - 1}| < |\beta'|(1 - \mu)$$

and this last term is $< (1 - \lambda)(1 - \mu)$. Therefore (3.8) is satisfied.

Finally, as (α, β) is a point in $\Omega_{\lambda,\mu}$, this set is not empty and (λ, μ) belongs to Θ by Proposition 3.17. □

The previous results and their proofs provide the following algorithm for generating any `SSM` matrix

$$A = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ h & g & f & e \\ d & c & b & a \end{pmatrix}.$$

with $a + d + g + f > 1$, $f > g$, and $b \neq c$.

Algorithm 3.20. (Generation of `SSM` matrices with given determinant.)
Input: K in $(0, 1)$.
Output: A strictly stochastic `SSM` matrix A with determinant K.

`Step 1:` Compute the unique positive root ν_0 of $r_K(z)$ following Remark 3.14.

`Step 2:` Take s randomly in $[\nu_0 + 1, 2)$.

`Step 3:` Take t randomly such that $|t| < \min\left\{2 - s, \sqrt{\frac{r_K(s-1)}{s-1}}\right\}$.

`Step 4:` Set $\lambda = \frac{s+t}{2}$ and $\mu = \frac{s-t}{2}$.

`Step 5:` Take $\alpha > 0$ randomly in $\left(\frac{\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu)}{\lambda}, \mu\right)$.

`Step 6:` Choose β randomly such that

$$\max\left\{-\lambda, \frac{\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu)}{\alpha}\right\} < \beta < \min\left\{\lambda, \frac{\frac{K}{\lambda+\mu-1} + (1-\lambda)(1-\mu)}{\alpha}\right\}.$$

`Step 7:` Choose β' randomly such that $\frac{|\alpha\beta - \frac{K}{\lambda+\mu-1}|}{1-\mu} < |\beta'| < 1 - \lambda$.

`Step 8:` Set $\alpha' := \frac{\alpha\beta - \frac{K}{\lambda+\mu-1}}{\beta'}$, $a := (\lambda + \beta)/2$,$b := (1 - \lambda - \beta')/2$,$c := (1 - \lambda + \beta')/2$ $d := (\lambda - \beta)/2$, $e := (1 - \mu - \alpha')/2$, $f := (\mu + \alpha)/2$, $g := (\mu - \alpha)/2$, and $h := (1 - \mu + \alpha')/2$.

`Final:` Return

$$A = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ h & g & f & e \\ d & c & b & a \end{pmatrix}.$$

Remark 3.21. As `SSM` matrices include `K81`* matrices, using Remark 3.10 we see that there exist matrices produced by the algorithm above that are not of type $\exp(Q)$.

Generating GMM matrices with a given determinant

For GMM matrices we do not have such a general result as in the previous sections. We do not know how to generate *any* strictly stochastic GMM matrix, but here we explain a way for generating some of them.

We could obtain a strictly stochastic matrix GMM matrix with determinant equal to K by exponentiating a rate matrix (i.e. a matrix with row sums equal to 0 and off-diagonal positive entries) with trace equal to $\log K$ (cf. (Pachter and Sturmfels, 2005a, Theorem 4.19)). However, not all GMM matrices are of this type (see Culver (1966) and Remark 3.10). We use that the product of two strictly stochastic matrices is again a strictly stochastic matrix in order to obtain a broader class of GMM matrices. In fact, we multiply a GMM matrix of type $\exp(Q)$ with determinant $\delta > K$ by a SSM matrix of determinant K/δ. We must admit that we do not know how much larger is this class of matrices. The set V of GMM matrices with determinant K corresponds to an affine variety of dimension 11. There are 11 free parameters for a rate matrix Q with given trace, so the matrices of type $\exp(Q)$ lie on a subset of V of dimension 11. Therefore the set of matrices produced by the algorithm below form a subset of maximum dimension of V, and this subset is larger than the set $\{\exp(Q)|Q$ rate matrix, $trQ = K\}$.

Algorithm 3.22. (Generation of GMM matrices with given determinant.)
Input: K in $(0, 1)$.
Output: A strictly stochastic GMM matrix A with determinant K.

```
Step 1: Take a random number $t$ in $(\log K, 0)$.
Step 2: Generate a random rate matrix $Q$ with nonzero entries and $trQ = t$.
Step 3: Compute $A_0 = \exp(Q)$.
Step 4: Following algorithm 3.20, generate a strictly stochastic SSM matrix $B$ with determinant equal to $K/e^t$.
 Final: Return $A = BA_0$.
```

The algorithms derived in this section have been implemented in C++ for practical use (see Chap. 7).

Chapter 4

Expectation-maxmization algorithm for parameter inference in equivariant models

Phylogenetic reconstruction focuses on inferring the phylogeny relating a set of taxa and estimating the evolutionary divergences between taxa. This is often done using a probabilistic evolutionary model and estimating the parameters that maximize the likelihood for the given data. There exist several effective methods for maximizing the likelihood under a continuous-time Markov process, and they are usually implemented when the rate matrix is fixed throughout the tree (homogeneous data). Here we consider the more general (discrete-time) Markov processes and we adapt the known Expectation-Maximization method to estimate the parameters of the transition matrices. We present the method for JC69*, K80*, K81*, SSM, and GMM evolutionary models and test it on simulated data. The results show a high performance in both transition matrices recovery and branch length estimation.

The inference of the parameters of the Markov process is often done by *maximum-likelihood estimation* (MLE): estimating the parameters that maximize the likelihood of observing given DNA sequences at the leaves of the tree.

The most widely used MLE methods, such as PAML Yang (2007), PHYLIP Felsenstein (1993), PAUP* Swofford (2003) are restricted to homogeneous continuous-time models such as Jukes-Cantor, Kimura two or three parameters, HKY or GTR.

There are two different approaches to estimate the parameters that maximize the likelihood for given data: one is to iteratively optimize the parameters for a given edge when the other parameters are fixed (Barry and Hartigan (1987), Jayaswal et al. (2011)), and the other is to globally optimize all parameters by estimating the hidden data. This later approach is known as *Expectation Maximization*(EM) and it was formally introduced by Dempster et al. (1977). EM has become a popular tool to deal with incomplete data problems or in problems which can be posed as such. That is to say, EM algorithm is used to compute the maximum likelihood estimate in the scenarios when the analytic solution to the likelihood equations cannot be obtained explicitly (e.g. missing data problems, models with latent variables, mixture or cluster learning) but the solution for the complete problem can be easily obtained. An exhaustive list of references and applications can be found in Tanner (1996), and more recently in McLachlan and Krishnan (2008). Here we present Empar, an MLE method based on the `EM` algorithm to estimate parameters of the (discrete-time) Markov evolutionary models.

Parameter estimates have strong impact on the branch length estimation (see e.g. Zou et al. (2011) and the references within). We test the proposed method on simulated data and we analyze the accuracy of the parameter and branch length estimate. We chose analog settings to Schwartz and Mueller (2010) for testing `Empar`. We evaluated it on four and six-taxon trees with several sets of branch lengths for different models and different alignment lengths. For the simulated data sets on these trees, we present an in-depth study of the performance of `Empar` and its dependence on factors such as model complexity, size of the tree, positioning of the branches, data and total tree lengths.

The algorithm works for any discrete-time models, for which the explicit form of the MLE can be given. We fix a set of n taxa. Let us recall that in accordance with the notation in the previous parts of the thesis, the set of nodes in $\mathcal{T}$ is denoted as $N(\mathcal{T})$, the set of leaves as $L(\mathcal{T})$, the set of interior nodes as $Int(\mathcal{T})$, and the set of edges as $E(\mathcal{T})$. We are given a set of DNA sequences associated to leaves of $\mathcal{T}$ and model of evolution along $\mathcal{T}$ as a discrete-time Markov process. We call $\pi = (\pi_{\mathtt{A}}, \pi_{\mathtt{C}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}})$ the distribution of nucleotides at the root r of $\mathcal{T}$ and $\theta = \{\pi, (A^e)_{e\in E(\mathcal{T})}\}$ the set of parameters for $\mathcal{T}$. Let X be a the set of 4^n possible patterns at the leaves of $\mathcal{T}$ and Y the set of $4^{|Int(\mathcal{T})|}$ possible patterns at the interior nodes of $\mathcal{T}$. Then the probability of observing nucleotides $\mathbf{x} = (x_l)_{l\in L(\mathcal{T})} \in X$ at the leaves of $\mathcal{T}$ and nucleotides $\mathbf{y} = (y_v)_{v\in Int(\mathcal{T})} \in Y$ at the interior nodes is

$$p_{\mathbf{x},\mathbf{y}}(\theta) = \pi_{y_r} \prod_{v\in Int(\mathcal{T})\setminus\{r\}} A^{e_{an(v),v}}_{y_{an(v)},y_v}$$

where $an(v)$ denotes the parent node of node v, $e_{an(v),v}$ is the edge for $an(v)$ to v, and $y_v = x_v$ if v is a leaf (cf. with the formula 1.1.

When the states at the interior nodes can be observed, this is called the *complete model*. However, in the usual situations the variables at the interior nodes are latent and then the probability of observing nucleotides $\mathbf{x} = (x_l)_{l\in L(\mathcal{T})}$ at the leaves of $\mathcal{T}$ under the *observed model* is

$$p_{\mathbf{x}}(\theta) = \sum_{\mathbf{y}=(y_v)_{v\in N(\mathcal{T})}\in Y} p_{\mathbf{x},\mathbf{y}}(\theta).$$

4.1 Expectation-Maximization algorithm

The data D we are given is a multiple sequence alignment and can be recorded into a vector of 4^n components $u_D = (u_{\mathbf{x}})_{\mathbf{x}\in X}$, where each $u_{\mathbf{x}}$ stands for the number of times pattern $\mathbf{x}$ appears as a column of the alignment. The likelihood function one wants to maximize is

$$\mathcal{L}_{obs}(\theta; u_D) = \prod_{\mathbf{x}\in X} p_{\mathbf{x}}(\theta)^{u_{\mathbf{x}}}.$$

Expectation maximization (EM) (Hartley (1958), Dempster et al. (1977)) was proposed as an attractive solution to obtaining maximum likelihood estimates (MLE's) when the formulas for the estimators are easy to obtain for a complete data model, but are rendered analytically intractable due to the incomplete data problem. If we have complete data cD observed at the interior nodes and leaves, we record it in an array $U_{cD} = (u_{\mathbf{x},\mathbf{y}})_{\mathbf{x}\in X,\mathbf{y}\in Y}$ where $u_{\mathbf{x},\mathbf{y}}$ is the number of times $\mathbf{x}$ was observed at the leaves and $\mathbf{y}$ at the interior nodes. The function to maximize for the complete data is

$$\mathcal{L}_c(\theta;U_{cD}) = \prod_{\mathbf{x}\in X,\mathbf{y}\in Y} p_{\mathbf{x},\mathbf{y}}(\theta)^{u_{\mathbf{x},\mathbf{y}}} = \prod_{\mathbf{x}\in X,\mathbf{y}\in Y} (\pi_{y_r} \prod_{v\in N(\mathcal{T})\setminus\{r\}} A^{e_{an(v),v}}_{y_{an(v)},y_v})^{u_{\mathbf{x},\mathbf{y}}}. \qquad (4.1)$$

As the complete model is a multinomial model, this likelihood function is guaranteed to have a global maximum which can be computed by an explicit formula. This formula must be given for each evolutionary model separately though. In the supplementary material we provide it for the SSM model (for the other models it can be obtained analogously).

EM algorithm is an iterative procedure alternating between the expectation (*E-step*) and maximization step (*M-step*). *E-step* uses the tree topology, the current estimates of model parameters and the observed data $u_D = (u_{\mathbf{x}})$ to assign a posterior probability to each of the possible $4^{|\mathtt{L}(\mathcal{T})|}$ patterns in X and give the most likely complete data u_{cD}. This step can be efficiently performed using the peeling algorithm of Felsenstein (2003). In the *M-step* the maximum likelihood estimates of the parameters are obtained by maximizing the likelihood of the complete model. Then one updates the parameters with these new estimates and iterates the process (see Fig. 4.1). The likelihood is guaranteed to increase at each iteration of this process (e.g. Wu (1983), Husmeier et al. (2005)) and, for a compact set of parameters, the algorithm converges to a critical point of the likelihood function. Although the output of the algorithm is not guaranteed to be a global maximum, multiple starting points are used to obtain optimality of the solution. An algebraic approach to the EM algorithm was introduced in Pachter and Sturmfels (2005b)[Chapter 12] and this encouraged us to apply it to the context of phylogenetic trees.

```
Require: $\mathcal{M}$- model, $\mathcal{T}$- phylogenetic tree, $u_D = (u_{\mathbf{x}})_{\mathbf{x}\in X}$ data vector.
  Initialize the values of the parameters $\theta$ such that $p_{\mathbf{x},\mathbf{y}}(\theta) > 0$ and choose a
  threshold $\epsilon > 0$.
  E-step: Define the expected complete data array $U = (u_{\mathbf{x},\mathbf{y}})_{\mathbf{x}\in X,\mathbf{y}\in Y}$:

      $u_{\mathbf{x},\mathbf{y}} := \frac{u_{\mathbf{x}}}{p_{\mathbf{x}}(\theta)} p_{\mathbf{x},\mathbf{y}}(\theta).$

  M-step: Compute the parameters $\theta^*$ that maximize the function (4.1) (including
  the root distribution).
  if $L_{obs;u_D}(\theta^*) - L_{obs}(\theta;u_D) > \epsilon$ then
    set $\theta := \theta^*$ and return to the E-step
  else
    $\hat{\theta} := \theta^*$
  end if
  return MLE $\hat{\theta}$ and likelihood of the observed model $\mathcal{L}_{obs}(\hat{\theta};u_D)$.
```

Figure 4.1: Expectation-Maximization algorithm for deriving the MLE estimates

4.2 Branch lengths

We recall the formula for the branch length of edge e (or of matrix A^e) given in (1.2):

$$l(A^e) = -\frac{1}{4}\log\det(A^e),$$

and denote the length of $\mathcal{T}$ by $\mathsf{L}_{\mathcal{T}} = \sum_{e\in|E(\mathcal{T})|} l(e)$.

Now we check that small errors in the estimates of the parameters ensure good recovery of the branch lengths. Let A and A' be two invertible 4×4 matrices such that $A - A'$ has small enough entries. Based on (1.2), we have

$$\begin{aligned} |l(A) - l(A')| &= \frac{1}{4}|\log\frac{\det(A)}{\det(A')}| = \frac{1}{4}|\log\det((A')^{-1}A)| \\ &= \frac{1}{4}|\log\det(\mathbf{Id} + (A')^{-1}(A - A'))| \\ &\approx \frac{1}{4}|\log(1 + Tr((A')^{-1}(A - A')))| \\ &\approx \frac{1}{4}|Tr((A')^{-1}(A - A'))| \leqslant \frac{1}{4}4||(A')^{-1}(A' - A))||_1 \\ &\leqslant ||(A')^{-1}||_1||A - A'||_1, \end{aligned} \tag{4.2}$$

where $||.||_1$ is the induced L_1 norm, defined as the maximum absolute column sum of a matrix (the approximations in the expression above hold if $(A')^{-1}(A - A')$ has small enough entries). Therefore if A' is a good approximation of A, the branch length computed from A' is also a good approximation of the branch length of A.

4.3 Maximum likelihood estimates

The *M-step* of the algorithm maximizes the likelihood of the complete model conditional on the current parameter estimates. Below we derive the MLE for the SSM model on a single branch e.

Let us index the letters $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ by $\{1, 2, 3, 4\}$. Following on the notation introduced before, $u_D = (u_{ij})_{i,j\in\{1,2,3,4\}}$ be the observed bases at the two end nodes of e. Let $\theta = \{\pi, A^e\}$ be the set of all parameters, where $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ is the root distribution. Recall that $\pi_1 = \pi_4, \pi_2 = \pi_3$. We deote by $A^e_{1,4}$ is the entry in the 1st row and the 4th column of A^e.

$$\begin{gathered} g_1(\theta) = A^e_{1,1} - A^e_{4,4}, \quad g_2(\theta) = A^e_{1,2} - A^e_{4,3}, \quad g_3(\theta) = A^e_{1,3} - A^e_{31}, \\ g_4(\theta) = A^e_{1,4} - A^e_{4,1}, \quad g_5(\theta) = A^e_{2,1} - A^e_{3,4}, \quad g_6(\theta) = A^e_{2,2} - A^e_{3,3}, \\ g_7(\theta) = A^e_{2,3} - A^e_{3,2}, \quad g_8(\theta) = A^e_{2,4} - A^e_{3,1}, \quad g_8(\theta) = A^e_{2,4} - A^e_{3,1}, \\ g_9(\theta) = 1 - A^e_{1,1} - A^e_{1,2} - A^e_{1,3} - A^e_{1,4}, \quad g_{10}(\theta) = 1 - A^e_{2,1} - A^e_{2,2} - A^e_{2,3} - A^e_{2,4}, \\ g_{11}(\theta) = 1 - A^e_{3,1} - A^e_{3,2} - A^e_{3,3} - A^e_{3,4}, \quad g_{12}(\theta) = 1 - A^e_{4,1} - A^e_{4,2} - A^e_{4,3} - A^e_{4,4}, \\ g_{13}(\theta) = \pi_1 - \pi_4, \quad g_{14}(\theta) = \pi_2 - \pi_3, \quad g_{15}(\theta) = 1 - \pi_1 - \pi_2 - \pi_3 - \pi_4. \end{gathered} \tag{4.3}$$

Taking the derivatives holds:

$$\begin{aligned}
\frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{1,1}} &= \frac{u_{1,1}}{A^e_{1,1}}+\lambda_1-\lambda_9, & \frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{1,2}} &= \frac{u_{1,2}}{A^e_{1,2}}+\lambda_2-\lambda_9,\\
\frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{1,3}} &= \frac{u_{1,3}}{A^e_{1,3}}+\lambda_3-\lambda_9, & \frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{1,4}} &= \frac{u_{1,4}}{A^e_{1,4}}+\lambda_4-\lambda_9,\\
\frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{2,1}} &= \frac{u_{2,1}}{A^e_{2,1}}+\lambda_5-\lambda_{10}, & \frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{2,2}} &= \frac{u_{2,2}}{A^e_{2,2}}+\lambda_6-\lambda_{10},\\
\frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{2,3}} &= \frac{u_{2,3}}{A^e_{2,3}}+\lambda_7-\lambda_{10}, & \frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{2,4}} &= \frac{u_{2,4}}{A^e_{2,4}}+\lambda_8-\lambda_{10},\\
\frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{3,1}} &= \frac{u_{3,1}}{A^e_{3,1}}-\lambda_8+\lambda_{11}, & \frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{3,2}} &= \frac{u_{3,2}}{A^e_{3,2}}-\lambda_7+\lambda_{11},\\
\frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{3,3}} &= \frac{u_{3,3}}{A^e_{3,3}}-\lambda_6+\lambda_{11}, & \frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{3,4}} &= \frac{u_{3,4}}{A^e_{3,4}}-\lambda_5+\lambda_{11},\\
\frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{4,1}} &= \frac{u_{4,1}}{A^e_{4,1}}-\lambda_4+\lambda_{12}, & \frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{4,2}} &= \frac{u_{4,2}}{A^e_{4,2}}-\lambda_3+\lambda_{12},\\
\frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{4,3}} &= \frac{u_{4,3}}{A^e_{4,3}}-\lambda_2+\lambda_{12}, & \frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial A^e_{4,4}} &= \frac{u_{4,4}}{A^e_{4,4}}-\lambda_1+\lambda_{12},\\
\frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial \pi_0} &= \frac{u_{0+}}{\pi_0}+\lambda_{13}-\lambda_{15}, & \frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial \pi_1} &= \frac{u_{1+}}{\pi_1}+\lambda_{14}-\lambda_{15},\\
\frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial \pi_2} &= \frac{u_{2+}}{\pi_2}-\lambda_{14}-\lambda_{15}, & \frac{\partial \mathcal{L}_{obs}(\theta,u_D)}{\partial \pi_3} &= \frac{u_{3+}}{\pi_3}-\lambda_{2,4}-\lambda_{15}.
\end{aligned} \tag{4.4}$$

Denote $x_1 = u_{1,1}+u_{4,4}$, $x_2 = u_{1,2}+u_{4,3}$, $x_3 = u_{1,3}+u_{4,2}$, $x_4 = u_{1,4}+u_{4,1}$, $x_5 = u_{2,1}+u_{3,4}$, $x_6 = u_{2,2}+u_{3,3}$, $x_7 = u_{2,3}+u_{3,2}$, $x_8 = u_{2,4}+u_{3,1}$.
Summing the sides of (4.4) gives

$$\begin{aligned}
\frac{1}{A^e_{00}}x_1 &= (\lambda_1-\lambda_9)+(-\lambda_1+\lambda_{12}) = \lambda_{12}-\lambda_9,\\
\frac{1}{A^e_{10}}x_5 &= (\lambda_5-\lambda_{10})+(-\lambda_5+\lambda_{11}) = \lambda_{11}-\lambda_{10},\\
\frac{1}{\pi_1}(u_{1+}+u_{4+}) &= \lambda_{13}-\lambda_{15}-\lambda_{13}-\lambda_{15} = -2\lambda_{15},\\
\frac{1}{\pi_1}(u_{2+}+u_{3+}) &= \lambda_{14}-\lambda_{15}-\lambda_{14}-\lambda_{15} = -2\lambda_{15}
\end{aligned}$$

and similarly

$$\begin{aligned}
\frac{1}{A^e_{1,2}}x_2 &= \frac{1}{A^e_{1,3}}x_3 = \frac{1}{A^e_{1,4}}x_4 = \lambda_{12}-\lambda_9,\\
\frac{1}{A^e_{2,2}}x_6 &= \frac{1}{A^e_{2,3}}x_7 = \frac{1}{A^e_{2,4}}x_8 = \lambda_{11}-\lambda_{20}.
\end{aligned}$$

Using the conditions of stochasticity:

$$\begin{aligned}
1 &= \sum_{i=1}^{4} A^{e}_{1,i} = \frac{\sum_{i=1}^{4} x_i}{\lambda_{12}-\lambda_9} = \frac{u_{1+}+u_{4+}}{\lambda_{12}-\lambda_9},\\
1 &= \sum_{i=1}^{4} A^{e}_{2,i} = \frac{\sum_{i=5}^{8} x_i}{\lambda_{11}-\lambda_{10}} = \frac{u_{2+}+u_{3+}}{\lambda_{11}-\lambda_{10}},\\
1 &= \sum_{i=1}^{4} \pi_i = \frac{2(u_{1+}+u_{2+}+u_{3+}+u_{4+})}{-2\lambda_{15}} = -\frac{u_{+}}{\lambda_{15}}.
\end{aligned}$$

As a result:

$$\begin{aligned}
\hat{A}^{e}_{1,1} = \hat{A}^{e}_{4,4} &= \frac{x_1}{u_{1+}+u_{4+}} = \frac{u_{11}+u_{44}}{u_{1+}+u_{4+}},\\
\hat{A}^{e}_{1,2} = \hat{A}^{e}_{4,3} &= \frac{x_1}{u_{1+}+u_{4+}} = \frac{u_{12}+u_{43}}{u_{1+}+u_{4+}},\\
\hat{A}^{e}_{1,3} = \hat{A}^{e}_{4,2} &= \frac{x_1}{u_{1+}+u_{4+}} = \frac{u_{13}+u_{42}}{u_{1+}+u_{4+}},\\
\hat{A}^{e}_{1,4} = \hat{A}^{e}_{4,1} &= \frac{x_1}{u_{1+}+u_{4+}} = \frac{u_{14}+u_{41}}{u_{1+}+u_{4+}},\\
\hat{A}^{e}_{2,1} = \hat{A}^{e}_{3,4} &= \frac{x_1}{u_{2+}+u_{3+}} = \frac{u_{21}+u_{34}}{u_{2+}+u_{3+}},\\
\hat{A}^{e}_{2,2} = \hat{A}^{e}_{3,3} &= \frac{x_1}{u_{2+}+u_{3+}} = \frac{u_{22}+u_{33}}{u_{2+}+u_{3+}},\\
\hat{A}^{e}_{2,3} = \hat{A}^{e}_{3,2} &= \frac{x_1}{u_{2+}+u_{3+}} = \frac{u_{23}+u_{32}}{u_{2+}+u_{3+}},\\
\hat{A}^{e}_{2,4} = \hat{A}^{e}_{3,1} &= \frac{x_1}{u_{2+}+u_{3+}} = \frac{u_{24}+u_{31}}{u_{2+}+u_{3+}},\\
\hat{\pi}_1 = \hat{\pi}_4 &= \frac{u_{1+}+u_{4+}}{u+},\\
\hat{\pi}_3 = \hat{\pi}_3 &= \frac{u_{2+}+u_{3+}}{u+}.
\end{aligned} \tag{4.5}$$

The MLE for the `JC69`*,`K80`* and the `K81`* models can be obtained analogously.

The algorithm was implemented for the `JC69`*, `K80`*, `K81`* and the `SSM` models. Chapter 8 is dedicated to testing its performance from a variety of angles. The algorithm was implemented in C++ and under the name `Empar` is available at `http://genome.crg.es/cgi-bin/phylo_mod_sel/AlgEmpar.pl`. Lastly, in chapter 10 we will apply the method to estimate branches of the species tree within different domains annotated in the framework of the *GENCODE* project.

Chapter 5

Background on phylogenetic varieties and equivariant models

Polynomials have always been present in statistical analyses as many models derived from the conditional independence models (e.g. polynomial regression).

Algebraic geometry studies the zero sets of polynomials. Methods taken from this or its sister fields of commutative algebra and combinatorics seem a natural support to study statistical models and aid their inference whenever a polynomial descriptions occur. The name *Algebraic Statistics* was coined by Pistone et al. (2000) in 2000. Up to that point, the application of algebra to statistics had been limited to a few specialized domains, e.g. experimental design, categorical data analysis and fixed and random effect linear models. Since then it has been a maturing discipline focused on the applications of algebraic geometry and its computational tools in the study of statistical models. Riccomagno (2008) gives a historical overview of the progress in the field since its conception. An extensive list of contributions to the field are given by Gibilisco et al. (2009).

Linear polynomials equations for models of contingency tables were used by Fienberg (1980). However, it was the seminal paper of Diaconis and Sturmfels (1998) that introduced the applicability of computational algebraic geometry in the context of exact test in the analyses of the contingency tables.

The field draws its tools not only from computational algebraic geometry but also from tropical, convex, and information geometry. More in-depth use of algebraic tools in experimental design was introduced in Pistone (1996). Kendall (1993) gives a brief survey of how computer algebra can be used in the implementation of the structures inherent to probability in statistics in order to aid the investigations in those fields.

As the field attracted scientists from a range of backgrounds, the spectrum of applications is broad. Graphical models are are an example of the field of study. From the algebraic perspective they can be described through the polynomials arising from the conditions on the variance-covariance matrix (Drton et al., 2007; Drton, 2008; Drton and Richardson, 2008). The dominant part of current research focuses on the Gaussian variables, however, the application of algebraic statistics to the field is not limited to discrete random variables. Other applications include model selection (Garcia-Puente, 2004) and the study of the properties of the maximum likelihood estimators– asymptotic properties of statistical models. e.g. shape of the likelihood function, the study of the regularity conditions or singularities, (Drton, 2009). Bayesian method are by no means an exception to the list of applications: e.g. in Bayesian networks (Garcia-Puente et al., 2005; Sullivant, 2008) and Bayesian model criterion (Drton and Foygel, 2008). The volume Drton et al. (2008) is an excellent collection of the recent advances.

The key observation is that the parameter spaces of certain statistical evolutionary models are semi-algebraic sets leads to the correspondence between the models and algebraic varieties. Algebraic versions of the evolutionary models have been introduced by Allman and Rhodes (2004b) and Pachter and Sturmfels (2005b). Drton and Sullivant (2007) give a following definition of a model in an algebraic setting.

Definition 5.1. *An algebraic statistical model* is a parametric statistical model, where the probability distribution is a polynomial function in the parameters.

Applications to the computational biology belongs the a young and fast-growing fields of interest. In particular, phylogenetics studies of evolutionary models and phylogenetics have been taken up by Allman and Rhodes (2003, 2004a, 2006a); Casanellas and Fernández-Sánchez (2007); Casanellas and Fernández-Sánchez (2008, 2011), to name a few. This includes novel tools of tree reconstruction– describing the genetic relationship between the species, and most recently model selection and their identifiability. This work, in particular chapter 6, is a contribution to the latter, extending the scope of the applications of algebraic statistics to phylogenetic mixtures models.

5.1 Background on algebraic geometry

Here we present basic concepts from algebraic geometry that we will use throughout the thesis. Recommended references for further reading are Cox et al. (2007), Hartshorne (1977), Harris (1992).

Affine varieties

Let $\mathtt{k}$ be a commutative field with unit and $\mathbb{A}^n_{\mathtt{k}}$ be the affine n-space over $\mathtt{k}$.

We will only consider $\mathtt{k}$ equal to the real numbers or the complex numbers $\mathbb{C}$, however, the results belowe hold in more generality. We will denote by $\mathtt{k}[\underline{x}]$, the polynomial ring with variables $\underline{x} = \{x_1, \ldots, x_n\}$.

Definition 5.2. *An affine algebraic variety $V \subset \mathbb{A}^k_n$* is the set of common zeroes of a collection of polynomials $S \subset \mathtt{k}[\underline{x}]$:

$$V = V(S) = \{x \in \mathtt{k}^n \mid f(x) = 0 \quad \forall f \in S\}$$

The empty set, the whole space $\mathbb{A}^n_k$, a finite union and an intersection of affine algebraic varieties are affine varieties. Consequently, affine algebraic varieties are the closed sets in what is called the *Zariski topology* in $\mathbb{A}^n_k$. Zariski closure of any set $Z \subseteq \mathbb{A}^n_k$, $\overline{Z}$, is defined as the smallest affine algebraic variety that contains it. Every non-empty Zariski open set in the affine space $\mathbb{A}^n_k$ is dense. In the remainder of this thesis we will naturally identify $\mathbb{A}^n_k$ with $\mathtt{k}^n$.

Example 5.3. *Linear subspaces of $\boldsymbol{k}^n$ are algebraic varieties. A single point in $\boldsymbol{k}^n$ is an algebraic variety: $(a_1, \ldots, a_n) = V(x_1 - a_1, \ldots, x_n - a_n), a_i \in \boldsymbol{k}$. As we will see inTheorem 5.7, the ideals of the form $< x_1 - a_1, \ldots, x_n - a_n >, a_i \in \boldsymbol{k}$ are exactly the maximal ideals of $\boldsymbol{k}[\underline{x}]$ if $\boldsymbol{k}$ is algebraically closed.*

Definition 5.4. An ideal $I \in \mathtt{k}[\underline{x}]$ is a subset of $\mathtt{k}[\underline{x}]$ satisfying:

1. $0 \in \mathcal{I}$,
2. if $f, g \in \mathcal{I}$, then $f + g \in \mathcal{I}$, and
3. if $f \in \mathcal{I}$ and $h \in \mathtt{k}[\underline{x}]$, then $hf \in \mathcal{I}$.

We say that an ideal $\mathcal{I}$ is *generated* by $f_1, \ldots, f_r$ if:

$$\mathcal{I} = \{\sum_{i=1}^{r} a_i f_i \mid a_i \in \mathtt{k}[\underline{x}]\}$$

In this case we will denote $\mathcal{I}$ by $(f_1, \ldots, f_r)$.

Definition 5.5. Let X be a subset $\mathtt{k}^n$. The *ideal* of X is defined as

$$\mathcal{I}(X) = \{f \in \mathtt{k}[\underline{x}] : f(x) = 0 \quad \forall x \in X\}.$$

Hilbert's basis theorem (Chap. 2 Cox et al., 2007) states that every ideal in $\mathtt{k}[\underline{x}]$ is finitely generated, i.e. for every ideal $\mathcal{I}$, there exists a finite set of polynomials $f_i \in \mathtt{k}[\underline{x}]$, s.t. $\mathcal{I} = (f_1, \ldots, f_s)$. In particular, any algebraic set $V(S)$ is an algebraic set for a finite collection of polynomials $V(S) = V(< S >) = V(f_1, \ldots, f_s)$.

Definition 5.6. The *radical* of an ideal $\mathcal{I}$ is defined as

$$\sqrt{\mathcal{I}} = \{f \in \mathtt{k}[\underline{x}] : f^n \in \mathcal{I} \text{ for some } n \geqslant 1\}.$$

The correspondence betwneen algebraic varieties and ideals is given by the following key theorem in algebraic geometry.

Theorem 5.7 (Hilbert's Nullstellensatz)**.** *Let $\boldsymbol{k}$ be an algebraically closed field. There is a $1-1$ correspondence between algebraic varieties in $\boldsymbol{k}^n$ and radical ideals in $\boldsymbol{k}[\underline{x}]$, given by $\mathcal{I}(V(J)) = \sqrt{J}$.*

Definition 5.8. A map $\Psi : \mathtt{k}^m \to \mathtt{k}^n$ is a regular map if

$$\Psi = (\Psi_1, \ldots, \Psi_n), \text{ with } \Psi_i \in \mathtt{k}[x_1, \ldots, x_m].$$

Definition 5.9. An algebraic variety $V \subseteq \mathtt{k}^n$ is a *cone* if for every $x \in V$, $\lambda x \in V, \forall \lambda \in \mathtt{k}$.

Projective varieties

Definition 5.10. The *projective space* of dimension n over $\mathtt{k}$, $\mathbb{P}^n_{\mathtt{k}}$, is defined as the set of equivalence classes in $\mathtt{k}^{n+1} \setminus \{0\}$ such that $(x_0, \dots, x_n) \sim (y_0, \dots, y_n)$ if there exists $\lambda \in \mathtt{k} \setminus 0$ for which $(x_0, \dots, x_n) = \lambda(y_0, \dots, y_n)$.

In geometric terms $\mathbb{P}^n_{\mathtt{k}}$ is often thought of as a set of lines through the origin in $\mathtt{k}^{n+1}$. Once a projective system has been chosen, homogeneous coordinates of a point in $x \in \mathbb{P}^n_{\mathtt{k}}$ are denoted by $[x_0 : \dots : x_n]$.

For convenience of the work of this thesis we will identify the affine space $\mathtt{k}^n$ with the subset $U = \{\underline{x} = [x_1, \dots, x_n] \in \mathbb{P}^n_{\mathtt{k}} \mid \sum x_i \neq 1\}$ of $\mathbb{P}^n_{\mathtt{k}}$,

Definition 5.11. A polynomial, $f \in k[x_0, \dots, x_n]$, is *homogeneous* if all its defining monomials have the same degree. In particular the degree of f is d if $f(\lambda x) = \lambda^d f(x)$, $\forall \lambda \in \mathtt{k}$. An ideal I in $k[x_0, \dots, x_n]$ is *homogeneous* if for all $f \in \mathcal{I}$ its homogeneous components are in $\mathcal{I}$. Alternatively, $\mathcal{I}$ is homogeneous if it is generated by homogeneous polynomials.

Definition 5.12. A *projective variety* is the zero set of a collection of homogeneous polynomials S:

$$\{x \in \mathbb{P}^n_{\mathtt{k}} : f(x) = 0 \text{ for all } f \in S\}.$$

Analogously to the affine case, we call $V(S)$ the projective variety defined by S.

As shown in Cox et al. (2007, Prop. 4, Chap. 8) $V(S)$ is well-defined: if $f(p) = 0$ for any set of homogeneous coordinates of $p \in \mathbb{P}^n_{\mathtt{k}}$, then $f(p) = 0$ for all homogeneous coordinates of p. As in the affine case we can also define a reverse process:

Definition 5.13. Let X be a subset of $\mathbb{P}^n$. The ideal of X is the set

$$\mathcal{I}(X) = \{f \in k[x_1, \dots, x_n] \mid f(p) = 0, \forall p \in X\}.$$

This is indeed an ideal and we see that it is homogeneous. The definition of a radical ideal translates into the projective setting. Moreover, as shown in Cox et al. (2007, Prop. 7, Chap. 3), the radical of a homogeneous ideal is itself a homogeneous ideal.

Theorem 5.14 (Thm 9, Chap. 8, p. 375, Cox et al. (2007), Projective strong Nullstellensatz)**.** Let $\mathtt{k}$ be an algebraically closed field, J a homogeneous ideal in $\mathtt{k}[\underline{x}]$ such that $\emptyset \neq V(J) \subseteq \mathbb{P}^n_{\mathtt{k}}$. We have that:

$$\mathcal{I}(V(J)) = \sqrt{J}.$$

As a result of the above theorem, in analogy to the affine case, there is a $1-1$ correspondence between proper radical homogeneous ideals and nonempty projective varieties (Cox et al., 2007, thm 10, Chap. 8, p. 375).

5.2 Algebraic evolutionary models

In section 1.4 we introduced basic notation in phylogenetics, including phylogenetic trees and evolutionary models (see Def. 1.3). Here we review these notions from the algebraic standpoint– we describe how to view hidden Markov processes on trees as algebraic varieties. We present a definition of a phylogenetic tree on a vector space, an algebraic presentation, parametrization and its associated algebraic variety, stochastic and projective algebraic varieties. These objects enable to use the lanugage of algebraic geometry in talking about phylogenetic objects and related problems and challenges in phylogenetics.

Let n a number and denote by $[n]$ the set $\{1, 2, \ldots, n\}$. For biological purposes, we think of $[n]$ as a set of sequences associated to certain taxa and we consider trees as connected acyclic graphs whose n leaves are bijectively labelled by the set $[n]$. Let $\mathcal{T}_n$ be the set of tree topologies (up to isomorphism) whose leaves are labelled by $[n]$. Trees in $\mathcal{T}_n$ are allowed to have any degree in its internal vertices. We recall that when the internal vertices of a tree $\mathcal{T} \in \mathcal{T}_n$ have degree 3, we say that the tree is *trivalent.*

We start by some definitions and notations required for subsequent chapters. We fix an ordered set $B = \{b_1, b_2, \ldots, b_k\}$ and we think of it as a basis of a $\mathbb{C}-$vector space $\mathcal{W} := \langle B \rangle_{\mathbb{C}}$. As mentioned in the previous section, in the applications to biology we take $B = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and think of its elements as nucleotides in a DNA sequence.

Below, we redefine a basic object in phylogenetics introduced in Defintion 1.3.

Definition 5.15. A *phylogenetic tree on* $\mathcal{W}$ is a tree $\mathcal{T}$ that has the vector space $\mathcal{W}_v := \mathcal{W}$ associated to each vertex v of $\mathcal{T}$. Usually the same notation $\mathcal{T}$ is used to represent both the graph and the phylogenetic tree. Elements of B at the vertices of $\mathcal{T}$ are thought as states of discrete random variables at the vertices.

Definition 5.16. Let $\mathcal{T}$ be a phylogenetic tree on $\mathcal{W}$ and assume that a distinguished vertex r of $\mathcal{T}$ (usually referred to as the *root*) is given, inducing therefore an orientation on all its edges. An *evolutionary presentation of* $\mathcal{T}$ is a vector $\pi = (\pi_{b_1}, \pi_{b_2}, \ldots, \pi_{b_k}) \in \mathcal{W}_r$, together with a collection of maps $\mathbf{A} = (A^{e_0,e_1})_{e \in E(\mathcal{T}), e=(e_0,e_1)}$ where each A^{e_0,e_1} belongs to $\mathrm{Hom}(\mathcal{W}_{e_0}, \mathcal{W}_{e_1})$.

From now on, we will identify vectors in $\mathcal{W}$ with its coordinates in the basis B written as a column vector. Similarly, we will identify the set $\mathrm{Hom}(\mathcal{W}, \mathcal{W})$ with the set of matrices with k rows and k columns and entries in the complex field by mapping any linear map to its matrix in the basis B. We take the convention that ta matrix$A = A^{e_0,e_1}$ in an evolutionary presentation act on $\mathcal{W}$ from the right (i.e. the action is $\omega^t \in \mathcal{W}_{e_0} \mapsto \omega^t A \in \mathcal{W}_{e_1}$). Recall that the vector $(1, 1, \ldots, 1) \in \mathcal{W}$ was denoted by $\mathbf{1}$.

Definition 5.17. An *algebraic evolutionary model* $\mathcal{M}$ is specified by giving a vector subspace $\mathcal{W}_0 \subset \mathcal{W}$ such that $\mathbf{1}^t \pi \neq 0$ for every $\pi \neq 0$ in $\mathcal{W}_0$, together with a multiplicatively closed subspace Mod of $\mathrm{Hom}(\mathcal{W}, \mathcal{W})$. A model is thus denoted by a pair: $\mathcal{M} = (\mathcal{W}_0, Mod)$. If $\mathcal{T}$ is a rooted phylogenetic tree on $\mathcal{W}$, then $\mathcal{T}$ *evolves under the algebraic evolutionary model* $\mathcal{M}$ if its evolutionary presentations lie in Mod and the vector π at the root belongs to $\mathcal{W}_0$. The set of evolutionary presentations of $\mathcal{T}$ that lie in $\mathcal{M}$ will be denoted by $\mathrm{Par}_{\mathcal{M}}(\mathcal{T}) = \mathcal{W}_0 \times \left(\prod_{e \in E(\mathcal{T})} Mod\right)$.

Remark 5.18. A subset of a ring is multiplcatively closed if for two elements that belong to it, so does their product: $x, y \in Mod$ implies that $xy \in Mod$. The reason for requiring this property in the subspace of the transition matrices is to ensure that if we multiply the matrices along an edge, the resulting matrix will remain in the model.

Remark 5.19. The condition $\mathbf{1}^t\pi \neq 0$ for every $\pi \in \mathcal{W}_0$ in the definition above means that, for a non-zero vector, the sum of the coordinates of the vectors in $\mathcal{W}_0$ is different from zero. The vectors in $\mathcal{W}_0$ represent the possible distributions for the root in the tree $\mathcal{T}$. The above condition is thus a plausible assumption for the models considered here and can be assumed without loss of generality and

Definition 5.20. Given a phylogenetic tree $\mathcal{T}$ on $\mathcal{W}$, $\mathcal{T} \in \mathcal{T}_n$, an $[n]$-*tensor* is any element of

$$\mathcal{L} := \otimes_{v \in [n]} \mathcal{W}_v = \otimes_{[n]} \mathcal{W}.$$

Notation 5.21. We will denote by $\mathcal{B} = B^n$ the set of n-words in B,

$$\mathcal{B} = \{\mathtt{X} = (\mathtt{x}_1, \ldots, \mathtt{x}_n) : \mathtt{x}_i \in B\}.$$

For the sake of simplicity in our notation, sometimes it will be convenient to identify every word $\mathtt{X} = (\mathtt{x}_1, \ldots, \mathtt{x}_n)$ with the tensor $\mathtt{x}_1 \otimes \ldots \otimes \mathtt{x}_n \in \mathcal{L}$ and consequently, we will identify $\mathcal{B}$ with the natural basis of $\mathcal{L}$. We will view a distribution $\mathbf{p} = (\mathbf{p}_{b_1 \ldots b_1}, \cdots, \mathbf{p}_{b_k \ldots b_k})$ on the set of elements in $\mathcal{B}$ at the leaves of a tree as the tensor in $\mathcal{L}$ having these coordinates in the basis $\mathcal{B}$, that is

$$\mathbf{p} = \sum_{\mathtt{x}_1, \ldots, \mathtt{x}_n \in B} \mathbf{p}_{\mathtt{x}_1 \ldots \mathtt{x}_n} \mathtt{x}_1 \otimes \ldots \otimes \mathtt{x}_n = \sum_{\mathtt{X} \in \mathcal{B}} \mathbf{p}_{\mathtt{X}}\, \mathtt{X}.$$

When an element in $\mathtt{X} = (\mathtt{x}_1, \ldots, \mathtt{x}_n) \in \mathcal{B}$ is used to refer to the coordinates corresponding to $\mathtt{x}_1 \otimes \ldots \otimes \mathtt{x}_n$ we will denote it by $\mathtt{x}_1 \ldots \mathtt{x}_n$. Therefore, $p \in \mathcal{L}$ can be represented as $\sum_{\mathtt{X} \in \mathcal{B}} p_{\mathtt{X}} \mathtt{X}$.

Definition 5.22. Given an algebraic evolutionary model $\mathcal{M}$, the *parametrization* of a rooted phylogenetic tree $\mathcal{T}$ on $\mathcal{W}$ evolving under the model $\mathcal{M}$ is the map

$$\Psi_{\mathcal{T}}^{\mathcal{M}} : \mathrm{Par}_{\mathcal{M}}(\mathcal{T}) \longrightarrow \mathcal{L} = \otimes_{[n]} \mathcal{W}$$

that corresponds to a hidden Markov process on the tree $\mathcal{T}$ when we restrict to stochastic matrices and distributions in $\mathcal{W}_0$. We recall that the leaves correspond to observed random variables and the interior nodes to hidden variables in the Markov process (see Section 1.4). That is, if the tree is rooted and directed from the root r, then the parameterization of $\mathcal{T}$ is the map

$$\Psi_{\mathcal{T}}^{\mathcal{M}}(\pi, \mathbf{A}) = \sum_{\mathtt{x}_1 \ldots \mathtt{x}_n \in \mathcal{B}} \mathbf{p}_{\mathtt{x}_1 \ldots \mathtt{x}_n} \mathtt{x}_1 \otimes \cdots \otimes \mathtt{x}_n$$

where $\mathbf{p}_{\mathtt{x}_1 \dots \mathtt{x}_n}$ is as established in 1.1:

$$\mathbf{p}_{\mathtt{x}_1 \dots \mathtt{x}_n} = \sum_{\mathtt{x}_v \in B, v \in Int(\mathcal{T})} \pi_{\mathtt{x}_r} \prod_{u \in N(\mathcal{T}) \setminus \{r\}} A^{e_{an(u)}, e_u}_{x_{an(u)}, x_u} \tag{5.1}$$

Here $\pi = (\pi_{\mathtt{x}})_{\mathtt{x} \in \mathcal{B}}$ are the coordinates in the basis $\mathcal{B}$ of the vector associated to the root.

From now on we will denote the coordinates of a point $\mathbf{p} \in \mathcal{L}$ in this basis as $\{\mathbf{p}_{x_1 \dots x_n}\}_{x_i \in B}$.

There are a few important properties of these parameterizations. Firstly, let us note that the position of the root plays a role in the above parameterization. However, under some assumptions its image is independent of it. The following lemma formalizes this idea. Let (u, v) be two adjacent vertices of an edge $\overline{e}$, and $\mathcal{T}_u$, $\mathcal{T}_v$ be the rooted versions of $\mathcal{T}$ on the two vertices, u and v, respectively (in these two trees the orientation of $\overline{e}$ is opposite).

Lemma 5.23 (Lemma 2.11, Casanellas et al. (2011))**.** *Let $\mathcal{T}_u$ be a rooted tree as above and consider an algebraic evolutionary model $\mathcal{M} = (\mathcal{W}_0, Mod)$. Let $(\pi, \mathbf{A})$ be an evolutionary presentation on $\mathcal{T}_u$ such that π has all its entries different from 0 and let $\tilde{\pi}^t = \pi^t A^{\overline{e}}$. Assume also that all the entries of $\tilde{\pi} \in \mathcal{W}_0$ are different from 0 and $D^{-1}_{\tilde{\pi}} (A^{\overline{e}})^t D_\pi$ belongs to Mod. Then, $\Psi^{\mathcal{M}}_{\mathcal{T}_u}(\pi, \mathbf{A}) = \Psi^{\mathcal{M}}_{\mathcal{T}_v}(\tilde{\pi}, \tilde{\mathbf{A}})$ if $\tilde{\mathbf{A}} = (\tilde{A}^e)_{e \in E(\mathcal{T}_v)}$, with*

$$\tilde{A}^e := \begin{cases} D^{-1}_{\tilde{\pi}} (A^{\overline{e}})^t D_\pi, & \textit{if } e = \overline{e}, \\ A^e, & \textit{otherwise} \end{cases}.$$

The models satisfying the conditions of the above lemma are called *root independent* (cf. Casanellas et al. (2011)).

Definition 5.24. We say that an algebraic evolutionary model $\mathcal{M} = (\mathcal{W}_0, Mod)$ is *root-independent* if it satisfies

1. $\tilde{\pi}^t := \pi^t A$ belongs to $\mathcal{W}_0$ for all $\pi \in \mathcal{W}_0$ and all $A \in Mod$, and
2. $D^{-1}_{\tilde{\pi}} (A^e)^t D_\pi \in Mod$ whenever $D^{-1}_{\tilde{\pi}}$ does exist.

The above lemma states that for root independent models the image of the parametrization map is independent of the position of the root. This leads to the non-identifiability issue for the placement of the root of a phylogenetic tree. Irrespective of the position of the root distribution, the joint probability $\mathbf{p}$ does not change, therefore we consider unrooted trees. We will prove lemma 5.23 in section 5.5 for a special subset of models, the so-called equivariant models, which include `JC69`*, `K80`*, `K81`*, `SSM` and `GMM`.

Definition 5.25. A *stochastic evolutionary model $s\mathcal{M}$* is specified by a subset $s\mathcal{W}_0$ of vectors in $\mathcal{W}$ whose entries sum to one, together with a multiplicatively closed set $sMod$ of complex matrices whose rows sum to one.

We want to point out that $s\mathcal{W}_0$ contains distributions (i.e. vectors with real and non-negative entries summing to 1) and $sMod$ contains stochastic matrices (i.e. matrices with real and positive entries and row sums equal to one). For a stochastic evolutionary model $s\mathcal{M}$, the space of matrices $sMod$ is not a vector substace anymore.

Example 5.26. If $\mathcal{M} = (\mathcal{W}_0, Mod)$ is an algebraic evolutionary model, define $s\mathcal{M} = (s\mathcal{W}_0, sMod)$ by taking $s\mathcal{W}_0 = \{\pi \in \mathcal{W}_0 : \mathbf{1}^t\pi = 1\}$ and $sMod = \{A \in Mod : A\mathbf{1} = \mathbf{1}\}$. Then, $s\mathcal{M}$ is a stochastic evolutionary model.

Definition 5.27. The *stochastic parametrization.* $\Psi_{\mathcal{T}}^{\mathcal{M}}$ of a rooted tree $\mathcal{T}$ evolving under a model $\mathcal{M}$ restricts to a polynomial map $\phi_{\mathcal{T}}^{\mathcal{M}}$ from

$$\mathrm{Par}_{s\mathcal{M}}(\mathcal{T}) = s\mathcal{W}_0 \times \left(\prod_{e \in E(\mathcal{T})} sMod \right)$$

to the hyperplane $H \subset \mathcal{L}$ defined by

$$H = \left\{ \mathbf{p} \in \mathcal{L} : \sum_{\mathtt{x}_1,\ldots,\mathtt{x}_n \in B} \mathbf{p}_{\mathtt{x}_1\ldots\mathtt{x}_n} = 1 \right\}.$$

To see why the image of the stochastic parameterization $\phi_{\mathcal{T}}^{\mathcal{M}}$ lies in H, we not that the map $\Psi_{\mathcal{T}}^{\mathcal{M}}$ restricted to distributions in $s\mathcal{W}_0$ and stochastic matrices in $sMod$ assigns to each set of parameters the corresponding distribution of patterns in $\mathcal{B}$ at the leaves of the tree. As a result, its image lies on the standard simplex in $\mathcal{L} = \otimes_{[n]}\mathcal{W}$ and, in particular, in the hyperplane H.

We proceed to define algebraic varieties associated to the parameterization maps.

Definition 5.28. The *affince phylogenetic variety* $CV_{\mathcal{T}}^{\mathcal{M}}$ associated to a phylogenetic tree $\mathcal{T}$ on $\mathcal{W}$ is

$$CV_{\mathcal{T}}^{\mathcal{M}} := \overline{\{\Psi_{\mathcal{T}}^{\mathcal{M}}(\pi_r, \mathbf{A}) : (\pi_r, \mathbf{A}) \in \mathrm{Par}_{\mathcal{M}}(\mathcal{T})\}}$$

where the closure is taken in the Zariski topology. Equivalently, $CV_{\mathcal{T}}^{\mathcal{M}}$ is the smallest algebraic set containing the image of $\Psi_{\mathcal{T}}^{M}$.

The *affine stochastic phylogenetic variety* $V_{\mathcal{T}}^{\mathcal{M}}$ associated to a phylogenetic tree $\mathcal{T}$ on $\mathcal{W}$ is

$$V_{\mathcal{T}}^{\mathcal{M}} := \overline{\{\phi_{\mathcal{T}}^{\mathcal{M}}(\pi_r, \mathbf{A}) : (\pi_r, \mathbf{A}) \in \mathrm{Par}_{s\mathcal{M}}(\mathcal{T})\}} \subset H$$

where the closure is taken in the Zariski topology.

There is a natural isomorphism between the points lying in the hyperplane $H = \{\mathbf{p} = (\mathbf{p}_{b_1\ldots b_1}, \ldots, \mathbf{p}_{b_k\ldots b_k}) \in \mathcal{L} : \sum \mathbf{p}_{\mathtt{x}_1\ldots\mathtt{x}_n} = 1\}$ and the open affine subset $\{\mathbf{p} = [\mathbf{p}_{b_1\ldots b_1} : \cdots : \mathbf{p}_{b_k\ldots b_k}] : \sum \mathbf{p}_{\mathtt{x}_1\ldots\mathtt{x}_n} \neq 0\}$ of $\mathbb{P}^{k^n-1} = \mathbb{P}(\mathcal{L})$ (we use projective coordinates $[\mathbf{p}_{b_1\ldots b_1} : \cdots : \mathbf{p}_{b_k\ldots b_k}]$ to distinguish them from affine coordinates, see Section 5.1). The *projective phylogenetic variety* $\mathbb{P}V_{\mathcal{T}}^{\mathcal{M}}$ associated to a phylogenetic tree $\mathcal{T}$ on $\mathcal{W}$ is the closure in $\mathbb{P}^{k^n-1} = \mathbb{P}(\mathcal{L})$ of the image of the stochastic parameterization $\phi_{\mathcal{T}}^{\mathcal{M}}$ defined above.

There is a close relation between the above varieties. As it is usually easier to deal with a homogeneous parameterization and homogeneous polynomials, it will be useful

.

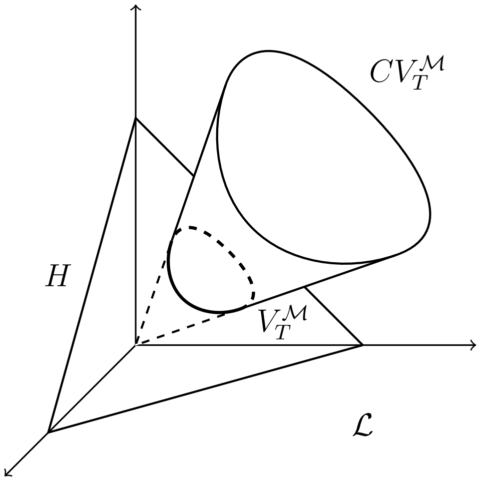

$CV_{\mathcal{T}}^{\mathcal{M}}$
H
$V_{\mathcal{T}}^{\mathcal{M}}$
$\mathcal{L}$

Figure 5.1: Affine $V_{\mathcal{T}}^{\mathcal{M}}$ and projective $CV_{\mathcal{T}}^{\mathcal{M}}$ phylogenetic varieties associated to a phylogenetic tree $\mathcal{T}$ on W (see Def. 5.28).

to prove that $CV_{\mathcal{T}}^{\mathcal{M}}$ is the cone over $\mathbb{P}V_{\mathcal{T}}^{\mathcal{M}}$. This is known for some particular models (for instance, see Allman and Rhodes (2008a) for a proof on the general Markov model) but as our definition of algebraic evolutionary model is quite general, we need to state it in its maximum generality.

Given a set $Z \subset \mathcal{L}$, we denote by $\mathcal{I}(Z)$ the ideal of polynomials in $\mathbb{C}[\mathcal{L}] := \mathbb{C}[p_{\mathtt{x}_1\dots\mathtt{x}_n}]$ that vanish over Z. Let us state a few facts relating the different phylogenetic varieties defined above for use in subsequent sections of this work.

Proposition 5.29. *Let $\mathcal{M} = (\mathcal{W}_0, Mod)$ be a root-independent evolutionary model and let $\mathcal{T}$ be a trivalent n-leaf tree on $\mathcal{W}$ evolving under $\mathcal{M}$ Then,*

(a) $CV_{\mathcal{T}}^{\mathcal{M}}$ equals the affine cone over the projective phylogenetic variety $\mathbb{P}V_{\mathcal{T}}^{\mathcal{M}}$;

(b) $\mathcal{I}(\operatorname{Im}\Psi_{\mathcal{T}}^{\mathcal{M}}) + (h) = \mathcal{I}(\operatorname{Im}\phi_{\mathcal{T}}^{\mathcal{M}})$, where $h = \sum \boldsymbol{p}_{\mathtt{x}_1,\dots,\mathtt{x}_n} - 1$;

(c) $V_{\mathcal{T}}^{\mathcal{M}} = CV_{\mathcal{T}}^{\mathcal{M}} \cap H$.

In particular, the polynomial equations defining $V_{\mathcal{T}}$ are formed by the homogeneous equations defining $CV_{\mathcal{T}}$ and with the extra stochastic equation $\sum \mathbf{p}_{\mathtt{x}_1\dots,\mathtt{x}_n} - 1 = 0$. In other words, the Corollary 5.29 states that $\dim CV_{\mathcal{T}}^{\mathcal{M}} = \dim \mathbb{P}V_{\mathcal{T}}^{\mathcal{M}} + 1$ and then if $\mathbf{p} = (\mathbf{p}_{b_1\dots b_1}, \dots, \mathbf{p}_{b_k\dots b_k})$ belongs to $CV_{\mathcal{T}}^{\mathcal{M}}$, then $q := [\mathbf{p}_{b_1\dots b_1} : \dots : \mathbf{p}_{b_k\dots b_k}]$ belongs to $\mathbb{P}V_{\mathcal{T}}^{\mathcal{M}}$. Moreover, if $s := \sum \mathbf{p}_{\mathtt{x}_1\dots\mathtt{x}_n} \neq 0$, then $q = [\frac{\mathbf{p}_{b_1\dots b_1}}{s} : \dots : \frac{\mathbf{p}_{b_k\dots b_k}}{s}]$ and $(\frac{\mathbf{p}_{b_1\dots b_1}}{s}, \dots, \frac{\mathbf{p}_{b_k\dots b_k}}{s})$ is a point in the affine stochastic phylogenetic variety $V_{\mathcal{T}}^{\mathcal{M}}$.

Consequence (a) was proved by Allman and Rhodes for the general Markov model (see (Allman and Rhodes, 2008a, Proposition 1)). As mentioned, the general proof can be found in Casanellas et al. (2011).

5.3 Use of algebraic geometry in phylogenetics

The use of algebraic geometry and its sister fields of computational algebra, commutative algebra and combinatorics, in phylogenetics is a failry new topic. Current trends are centered around describing phylogenetic objects through polynomial equations. Using the algebraic techniques in phylogenetic inference falls under the umbrella of algebraic statistics. Inference in phylogenetics include finding the discrete (underlying tree topology) and estimating the continuous parameters (parameters of the evolutionary models).

Invariants were introduced in Section 1.4.3– these are algebraic relations that are satisfied by the joint probability distribution under a given evolutionary model. More formally, we have the following definitions.

Definition 5.30. Let $\mathcal{T}$ be an $n-$taxon tree, $\mathcal{M}$ a model and $V_{\mathcal{T}}^{\mathcal{M}}$ its associated affine stochastic phylogenetic variety in $\mathbb{A}^{4^n}$. An *invariant* is a polynomial in the ideal $\mathcal{I}(V_{\mathcal{T}}^{\mathcal{M}})$. A *phylogenetic invariant* is an element in $\mathcal{I}(V_{\mathcal{T}}^{\mathcal{M}})$ for $\mathcal{T}$, but not in $\mathcal{I}(V_{\mathcal{T}'}^{\mathcal{M}})$ for some other $\mathcal{T}'$ i.e. not in $\bigcap_{\mathcal{T}'} \mathcal{I}(V_{\mathcal{T}'}^{\mathcal{M}})$. A *model invariant* is an element in the intersection of $\mathcal{I}(V_{\mathcal{T}}^{\mathcal{M}})$ for all the tree topologies $\mathcal{T}$ on n-taxa.

Phylogenetic invariants are beyond the scope of this thesis. We are interested in model invariants, i.e, generators of the ideal of the model that vanish on all tree topologies. Here we give a few examples of the computation of invariants.

Example 5.31. Consider the claw tree $\mathcal{T}$ on $n = 3$ labeled leaves $\{X_1, X_2, X_3\}$ and a $\mathcal{M} = \texttt{GMM}$ model on two sates $\{0, 1\}$ (cf. Example 1.9). We write a parameterization map:

$$\phi_{\mathcal{T}}^{\mathcal{M}} : \mathrm{Par}_s^{\texttt{GMM}}(\mathcal{T}) \longrightarrow \mathcal{L} = \{p_{\texttt{AAA}}, ..., p_{\texttt{TTT}}\}$$

that corresponds to a hidden Markov process on $\mathcal{T}$. It takes the stochastic parameters to the joint probabilties.

Here we show a code in Singular (Greuel et al., 2001) that can be used to compute the ideal of $V_{\mathcal{T}}^{\mathcal{M}}$.

```
int b=2;
ring r1 = 0,(p(1..b)(1..b)(1..b)),dp;
ring r2 = 0,(m1(1..b)(1..b),m2(1..b)(1..b),m3(1..b)(1..b),r(1..b)),dp;


int i,j,k,l,s;
poly p, p1,p2,p3;
list L,Ls;
s = 1;
for (i=1; i<=b; i=i+1)
{
  for(j=1; j<=b; j=j+1)
  {
    for(k=1; k<=b; k=k+1)
    {
      p=0;
      for(l=1; l<=b; l=l+1)
      {
        p = p + r(l)*m1(l)(i)*m2(l)(j)*m3(l)(k);
      }
      L[s] = p;
      s = s+1;
    }
  }
}

p=0;
for(j=1; j<=b; j=j+1)
{
  p = p + r(j);
}

Ls[1] = p - 1;
s=2;
for(i=1; i<=b; i=i+1)
{
  p1=0; p2=0; p3=0;
  for(j=1; j<=b; j=j+1)
  {
    p1 = p1 + m1(i)(j);
    p2 = p2 + m2(i)(j);
    p3 = p3 + m3(i)(j);
  }
  Ls[s] = p1 - 1;
  Ls[s+1] = p2 - 1;
  Ls[s+2] = p3 - 1;
  s = s+3;
}

map f=r1,L[1..b^3];
ideal I0=0;
setring r1;
ideal J=preimage(r2,f,I0);
print(J);

dim(std(J));
```

The output is of the code is:

```
p(1)(1)(1)+p(1)(1)(2)+p(1)(2)(1)+p(1)(2)(2)
  +p(2)(1)(1)+p(2)(1)(2)+p(2)(2)(1)+p(2)(2)(2)-1
7
```

We see that the only polynomial vanishing in the image of the parametrization corresponds to the stochastic condition and thus the image fills the whole space. This is also given by the dimension of $V_T^{\mathcal{M}}$, which is 7. If we replace the ideal of stochastic conditions coded in line "ideal $I0 = Ls[1..3 * b + 1]$" by an empty ideal, $I0 = 0$, we obtain the ideal of $CV_T^{\mathcal{M}}$. We were not able to compute the above example for $b = 4$ (e.g. $B = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$)– the computations did not finish within days.

We might be interested in computing the linear part of a generating ideal of $CV_T^{\mathcal{M}}$. As we will see in chapter 9 it is precisely the linear part this ideal that is of interest in phylogenetic model selection.

We give two examples of this computation performed in *Singular*. There are two functions that make this computation possible: `degBound` works only for homogeneous ideals and limits the degree in the computations of Grobner basis, i.e. $degBound = 5$ produces a basis up to degree 5; and `nselect`, take an ideal as an input and keeps the polynomials which do not contain variables in the prespecified range, i.e. $nselect(\mathcal{I}, 1..84)$ keeps the polynomials of the ideal $\mathcal{I}$ that are not expressed the first 84 indeterminates.

Example 5.32. [`GMM`] Let us consider unrooted trees on the set of $\{1, 2, 3, 4\}$ leaves and denote them by $12 - 34, 14 - 23, 13 - 24, \tau_4$, where τ_4 is a star tree (eg. $12 - 34$ has pairs (12) and (34) in separate clades joined by an internal edge). Consider the `GMM` model on these trees and the corresponding parameterizations and calculate the linear part of the ideal $CV_T^{\mathtt{GMM}}$. In fact, it is known that the only linear invariant of $V_T^{\mathtt{GMM}}$ is the stochastic condition, so $CV_T^{\mathtt{GMM}}$ will have no linear invariants. The *Singular* code for this example is given in the appendix A.

It is known that phylogenetic invariants for $12|34$ are the 5×5 minors of the flattening of the joint vector $(p_{\mathtt{AAAA}}, \dots, p_{\mathtt{TTTT}})$ along the bipartition $12|34$ vanish on the phylogenetic variety (and respectively for other tree topologies and their partitions).

Example 5.33. Let us take τ_3 to be a 3-leaf star tree. Using the same apprach as in the example above, we can calculate the linear part of $\mathcal{I}(V_{\tau_3}^{\mathtt{ATR}})$. The code is provided in the appendix A.

Needless to say, the above procedure cannot be performed in a reasonable time for larger trees. Alternative approaches are needed. In subsequent sections we will see an example of such– we propse a novel approach to computing all model invariants via an algorithm based on group theory. It gives a faster and method to obtaining the sets of invariants for fairly large trees. Most importantly, it sheds light on the behaviour of these invariants and proves the intution that the model invariants are valid for phylogenetic mixture models.

5.4 Groups and actions

Tools from representation theory lie at the basis of the methods developed in chapter 6, where we describe the linear structure of the phylogenetic mixtures and their dimension. In particular, we will connect subgroups of a general symmetric group on B to phylogenetic evolutionary models. More on group theory can be found in Rotman (1995) and for background on linear represnetation of groups we refer to Serre (1977).

Elementaries. We restrict to finite groups and we use multiplicative notation for the group operation. Let $|G|$ denote the *order* (cardinality) of a group G. A mapping $\psi : G \to H$ between two groups preserving the group structure, that is $\psi(g_1)\psi(g_2) = \psi(g_1 g_2)$ for any $g_1 g_2 \in G$ is called a *homomorphism.* A one-to-one (bijective) homomorphism is an *isomorphism.*

Example 5.34. The *symmetric group* on a set of cardinality k, $\mathfrak{S}_k$, is the set of all permutations of k elements. We have that $|\mathfrak{S}_k| = k!$. By definition, $\mathfrak{S}_k$ contains the identity element and the inverses of all its elements. The *dihedral group* on k, D_k, elements is a group of symmetries of a k-sided regular polygon. We have that $|D_k| = 2k$.

If a subset H of G is a group under the group operation of G, then H is called a subgroup of G denoted by $H \leqslant G$. Any group is its own subgroup, and $\{id\}$ is a subgroup of any group. If $H \neq G$, then H is a proper subgroup $H < G$.

Definition 5.35. If $H < G$ and $g \in G$, then $Hg = \{hg : h \in H\}$ is a *right coset* of H in G. Any element of Hg (including g) is called a *representative* of Hg. Any two right cosets are either disjoint or equal and we write $H \setminus G$ for the *quotient space* of right cosets

$$H \setminus G = \{Hg : g \in G\}.$$

Remark 5.36. There is a corresponding definition of a left coset, however, in thi thesis we will only require the right cosets.

Definition 5.37. The *index* of H in G, $[G : H]$, id the number of right cosets of H in G.

Theorem 5.38 (Lagrange's theorem)**.** The order of any subgroup H of a finite group G divides the order of the group. Moreover, the following equality holds:

$$[G : H] = \frac{\mid G \mid}{\mid H \mid}. \tag{5.2}$$

Remark 5.39. Therefore, the quotient in the equation (5.2) is an integer.

Definition 5.40. A subset S of G is called a *transversal* for $H \setminus G$ if for any distinct elements $g_1, g_2 \in S$, $Hg_1 \neq Hg_2$ and G can be partition in the following way:

$$G = \bigcup_{i \in S} Hg_i. \tag{5.3}$$

In other words, the transversal is the set of coset representatives.

The cardinality of a transversal for $H \setminus G$ if therefore $[G : H]$.

Definition 5.41. An *action* of a group G on a set Y is a map $G \times Y \to Y$denoted by $(g, y) \mapsto gy$, $g \in G, y \in Y$, such that $g_1(g_2y) = (g_1g_2)y$, $idy = y$. We also say that G *acts* on Y. If G acts on Y then for any $y \in Y$ the *stabilizer (isotropy subgroup)* of y is defined as:

$$G_y = \{g \in G : gy = y\}.$$

The *orbit* of y is

$$\{y\}_G = \{gy : g \in G\}.$$

Remark 5.42. Similarly, if $X \in \mathcal{B}$, we denote by G_X the stabilizer of X: $G_X = \bigcap_{i=1}^{n} G_{x_i}$, and we write $X_G = \{gX : g \in G\}$ for the orbit of X. Note that if $X \in \mathcal{B}$, then $gX \in \mathcal{B}$ for every $g \in G$.

Lemma 5.43. Let $H \leqslant G$ and G act on a set Y and $\{g_1, \ldots, g_m\}$ be a transversal for $H \setminus G$. For every $y \in Y$, we have

$$\{y\}_G = \bigcup_{i=1,\ldots,m} \{g_iy\}_H.$$

Proof. We apply the decomposition (6.4) to an element y:

$$\{y\}_G = \{gy : g \in G\} = \{hg_iy : h \in H, g_i \in S\} = \bigcup_{g_i \in S} \{h(g_iy) : h \in H\} = \bigcup_{g_i \in S} \{g_iy\}_H$$

□

Theorem 5.44 (orbit-stabilizing theorem)**.** *Let G be a group acting on a set Y and let $y \in Y$, G_y be the stabilizer and $\{y\}_G$ the orbit of y. There exists a bijection*

$$\{y\}_G \cong G/G_y.$$

In particular, by Lagrange's theorem it follows that:

$$\mid \{y\}_G \mid = \frac{\mid G \mid}{\mid G_y \mid}$$

Representation theory Let G be a finite group and let V be a $\mathbb{C}-$vector space of finite dimension. Let $GL(V)$ be the group of isomorphisms of V onto itself- *the general linear group of* V. By choosing a basis of V an isomorphism $a \in GL(V)$, $a : V \longrightarrow V$ can be identified with an invertible square matrix.

Definition 5.45. A (linear) *representation* of G in V is a group homomorphism $\rho : G \to GL(V)$.

V is called the *representation* space and its dimension is the dimension of the representation. We will refer to the representation as ρ or V depending on the context.

For notational convenience we will use ρ_g and $\rho(g)$ interchangeably. The representation defines an action on V as $(g, v) \mapsto P_g(v)$ so that $P_g(v)$ will be also denoted by $g\dot{\mathbf{x}}$ (in the sense of the Definition 5.41) and V will be understood as a $G-$module.

Definition 5.46. A vector space $V' \subset V$ is *stable* (invariant) under the action of G if $\rho_g(v') \in V'$ for all $g \in G, v' \in V'$.

Definition 5.47. A function $f \in \mathbb{C}^G$ is a *class function* if for all $g_1 g_2 \in G$ $f(g_1) = f(g_2 g_1 g_2^{-1})$.

It is possible to summarize the information about the representation in a compact form through the notion of character.

Definition 5.48. The character of a representation ρ of G is the function $\chi_\rho : G \to \mathbb{C}$ defined by:

$$\chi_\rho(g) = Tr(P_g).$$

Two representations with the same character are isomorphic (see Serre (1977, Chap. 2.3, Cor.2)). The character of an irreducible representation is called an *irreducible character.*

Definition 5.49. We say that a representation $\rho : G \to GL(V)$ is *irreducible* if it is not 0 and no $W \subset V$, except for 0 and V, is stable under G.

Definition 5.50. Given an element $g \in G$, the *conjugacy class* of g is defined as $C(g) = \{h^{-1}gh : h \in G\}$.

Being in the same conjugacy class is an equivalence relation that partitions G into non-overlapping sets: if $g_1, g_2 \in G$, we have that either $C(g_1) = C(g_2)$ or $C(g_1) \cap C(g_2) = \emptyset$. If $C_1, \ldots, C_s$ are the conjugacy classes for G, write $\mathcal{C}(G)$=($\mid C_1 \mid, \ldots, \mid C_s \mid$) for the s-tuple of their cardinalities, so that $\sum_{i=1}^{s} \mid C_i \mid = |G|$.

From the definition (5.47) it is clear that a character is a class function on G: $\chi(g_1) = \chi(g_2)$ whenever $\mathcal{C}(g_1) = \mathcal{C}(g_2)$. We can write: $\chi(G) = (\chi(\mathcal{C}_1), \ldots, \chi(\mathcal{C}_s))$.

Definition 5.51. *A character table* is a 2-way table, where the columns are labeled by a set of representatives of conjugacy classes and the rows are labelled by the irreducible characters. The entries are the irreducible characters evaluated on a given conjugacy class.

Definition 5.52. Let V, V' be a two representation of G. The direct sum $V \oplus V'$ of V and V' is also a representation with the action of the group given by $\rho_{V\oplus V'}(v, v') = (\rho_V(v), \rho_{V'}(v'))$. The tensor product of V and V', $V \otimes V'$, is again a representation with the action of the group given by $\rho_{V\otimes V'}(v \otimes v') = \rho_V(v) \otimes \rho_{V'}(v')$.

Remark 5.53. The above constructions can be generalized to finite sums and products.

The character of a direct sum of representations is the sum: $\chi_{V\oplus V'}(g) = \chi_V(g) + \chi_{V'}(g)$. The character of a tensor product of representations is the product of characters: $\chi_{V\otimes V'}(g) = \chi_V(g)\chi_{V'}(g)$.

Theorem 5.54 (Serre (1977), Chap. 2.5, Thm 6)**.** The set $\Omega = \{\omega_i\}_{i=1,\ldots,t}$ of irreducible characters of a group G forms an orthonormal basis of the class functions relative to the inner product defined by

$$\langle f, h\rangle = \frac{1}{|G|}\sum_{g\in G} f(g)\overline{h(g)}. \tag{5.4}$$

Theorem 5.55 (Maschke's Theorem)**.** Let $\Omega = \{\omega_i\}_{i=1,\ldots,t}$ be the set of irreducible characters of G. For every linear representation V there exists a decomposition of V into its *isotypical* components:

$$V = \oplus_{i=1}^{s} V[\omega_i], \tag{5.5}$$

where each $V[\omega_i]$ is isomorphic to a number of copies of a irreducible representation N_i associated to ω_i, $V[\omega_i] \cong N_i \otimes \mathbb{C}^{m_i}$ for some positive integer m_i, called the *multiplicity of* V *relative to* ω_i. Moreover, if ρ is the character of V, then $m_i = \langle \rho, w_i\rangle$

As a consequence of the above, the number of irreducible characters equals the number of conjugacy classes of G.

Since $\chi_{\otimes^n V} = \chi_W^n$, if w_i isthe irreducible character of a group G

$$m_1(n) = \frac{1}{|G|}\sum_{i=1}^{s} \chi^n(\mathcal{C}_i) \mid \mathcal{C}_i \mid, \tag{5.6}$$

Definition 5.56. Let G be a subgroup of the symmetric group $\mathfrak{S}_k$ of a set B of k elements. The symmetric group $\mathfrak{S}_k$ acts naturally on B and if W is the $\mathbb{C}-$vector space $\langle B\rangle_{\mathbb{C}}$ it gives rise to a linear representation:

$$\rho : \mathfrak{S}_k \to GL(W),$$

$$\sigma \mapsto P_\sigma,$$

where P_σ is linear map defined by permuting elements in B according to σ. This is called the *defining representation* of $\mathfrak{S}_k$.

Any subgroup $G < \mathfrak{S}_k$ acts also on B and the defining representation ρ restricts to a representation of G. This restriction of ρ to G will be also called the *defining representation* of G.

The following definition and lemma will be used in Section 6.1, here they are given in a general form.

G acts in $\mathcal{B}$ in the following way, if $X = (x_1 \ldots x_n)$

$$g(X) = g(x_1)\ldots g(x_n)$$

and gives a representation in $\otimes_n W$ as specified in (5.52).

Definition 5.57. Given a set of taxa n, a G*-tensor* on n is an n-tensor invariant by the action defined in (5.52). The set of G-tensors will be denoted by $\mathcal{L}^G$.

We have that $\chi = \sum_{i=1}^{s} < \chi, \omega_i > \omega_i$ and the dimension of the space of invariants equals the number of trivial representations in the decomposition:

$$m_1 =< \chi, \omega_1 >= \frac{1}{|G|} \sum_{g \in G} \chi(g)\overline{\omega_1} = \frac{1}{|G|} \sum_{i=1}^{s} \chi(\mathcal{C}_i)\overline{\omega_1(\mathcal{C}_i)} \mid \mathcal{C}_i \mid . \tag{5.7}$$

The following lemma states that the set of stable elements by the action of a group G (in the sense of the definition 5.46) can be obtained from systems of linear equations associated with a system of generators of G.

Lemma 5.58. *Let V be the set of elements of a vector space invariant by the action of $G = \langle g_1, \ldots, g_t \rangle$. We have that*

$$V^G = \bigcap_{i=1}^{t} V^{\langle g_i \rangle}.$$

Proof. The $(\rightarrow)$ inclusion is straightforward. To prove the second inclusion, let $p \in \bigcap_{i=1}^{s} V^{\langle g_i \rangle}$, so we have $g_i p = p$ for any i. Let $g \in G$ be any element of the group and write $g = g_{i_1}^{m_1} \ldots g_{i_r}^{m_t}$ with $m_i > 0$. Adopting the convention of the right to the left action of the group elements, the recursive application of the g_i to p completes the proof: $gp = g_{i_1}^{m_1}(g_{i_2}^{m_2} \ldots g_{i_r}^{m_r} p) = p$. □

5.5 Equivariant models of evolution

In this section we study in a mathematical setting a Markov models of evolution introduced in Section 1.4.

Let B be a set of k elements and $\mathcal{W} = \langle B \rangle_{\mathbb{C}}$.

Definition 5.59. Let G be a permutation group of B (that is, a group whose elements are permutations of the set B, $G \leqslant \mathfrak{S}_k$). Given $g \in G$, write P_g for the $k \times k$-permutation matrix corresponding to g: $(P_g)_{i,j} = 1$ if $g(j) = i$ and 0 otherwise. The *G-equivariant evolutionary model*, $\mathcal{M}_G$, is defined by taking Mod equal to $\mathrm{Hom}_G(\mathcal{W}, \mathcal{W})$, that is,

$$\mathrm{Hom}_G(\mathcal{W}, \mathcal{W}) = \{A \in M_{k,k}(\mathbb{C}) \mid AP_g = P_g A, \forall g \in G\}$$

and $\mathcal{W}_0 = \{\pi \in W \mid P_g \pi = \pi \, \forall g \in G\}$. It is clear that the above subsets define vector subspaces of $\mathrm{Hom}(\mathcal{W}, \mathcal{W})$ and $\mathcal{W}$. On the other hand, if $A_1, A_2 \in \mathrm{Hom}_G(\mathcal{W}, \mathcal{W})$, then

$$P_g A_1 A_2 P_g^{-1} = (P_g A_1 P_g^{-1})(P_g A_2 P_g^{-1}) = A_1 A_2$$

so that $A_1 A_2 \in \mathrm{Hom}_G(\mathcal{W}, \mathcal{W})$. Therefore, equivariant models are examples of algebraic evolutionary models in the sense of Definition 5.17.

Below we view some of the models introduced in Section 1.4 as equivariant models via its associated subgroup of symmetries and give their characteristics using notions from Section 5.4.

Notation 5.60. For an equivariant model $\mathcal{M}$ we denote by $G_{\mathcal{M}}$ its corresponding group. We call ρ the defining representation of $G_{\mathcal{M}}$ (see Def. 5.56) and we denote by $\chi_{\mathcal{M}}$ its character. From now on B will be the set $B = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and χ^n will denote the character of the defining presentation.

Jukes-Cantor (JC69*). A transition matrix of the JC69* model has the form given by 1.11. In order to view it as an equivariant model we observe that its substitution matrices are invariant under any permutation of rows and columns.

The associated group is the symmetric group $G_{\mathtt{JC69}^*}$=$\mathfrak{S}_4$. Its cardinality is 24 and its the elements correspond to all permutations of 4 letters (see Defintition 5.34):

$$\begin{aligned} G_{\mathtt{JC69}^*} = \{&id, (\mathtt{AG}), (\mathtt{AC}), (\mathtt{AT}), (\mathtt{CG}), (\mathtt{CT}), (\mathtt{GT}), \\ &(\mathtt{CGT}), (\mathtt{ATG}), (\mathtt{ACT}), (\mathtt{AGC}), (\mathtt{AGT}), (\mathtt{ATC}), (\mathtt{ACG}), (\mathtt{CTG}), \\ &(\mathtt{AC})(\mathtt{GT}), (\mathtt{AG})(\mathtt{CT}), (\mathtt{AT})(\mathtt{CG}), (\mathtt{CT})(\mathtt{AG}), \\ &(\mathtt{ACGT}), (\mathtt{ATGC}), (\mathtt{AGCT}), (\mathtt{ATCG}), (\mathtt{ACTG}), (\mathtt{AGTC})\}. \end{aligned} \tag{5.8}$$

$G_{\mathtt{JC69}^*}$ can be generated by 2 elements consiting of a transposition and a cycle, e.g. $G_{\mathtt{JC69}^*} =< (\mathtt{AC}), (\mathtt{ACGT}) >$.

Example 5.61. The defining representation of $G_{\mathtt{JC69}^*}$ applied to its generators is:

$$(\mathtt{AC}) \mapsto P_g^1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, (\mathtt{ACGT}) \mapsto P_g^2 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

It is clear that $P_g^1, P_g^2 \in \mathrm{Hom}_{\mathtt{JC69}^*}(\mathcal{W}, \mathcal{W})$.

A set of representatives of the conjugacy classes can be given by

$$\{id, (\mathtt{AC})(\mathtt{GT}), (\mathtt{ACGT}), (\mathtt{AG}), (\mathtt{ACG})\}.$$

Indexing the conjugacy classes in the order of these representants we have:

$$\begin{aligned} \mathcal{C}_1 &= \{id\}, \mathcal{C}_2 = \{(\mathtt{AC})(\mathtt{GT}), (\mathtt{AG})(\mathtt{CT}), (\mathtt{AT})(\mathtt{CG})\}, \\ \mathcal{C}_3 &= \{(\mathtt{ACGT}), (\mathtt{ATGC}), (\mathtt{AGCT}), (\mathtt{ATCG}), (\mathtt{ACTG}), (\mathtt{AGTC})\}, \\ \mathcal{C}_4 &= \{(\mathtt{ATG}), (\mathtt{ACT}), (\mathtt{AGC}), (\mathtt{AGT}), (\mathtt{ATC}), (\mathtt{ACG}), (\mathtt{CGT}), (\mathtt{CTG})\}, \\ \mathcal{C}_5 &= \{(\mathtt{AG}), (\mathtt{AC}), (\mathtt{AT}), (\mathtt{CG}), (\mathtt{CT}), (\mathtt{GT})\}. \end{aligned}$$

Therefore the transversal set can be given by

$$\{id, (\mathtt{AC})(\mathtt{GT}), (\mathtt{ACGT}), (\mathtt{ACG}), (\mathtt{AC})\}.$$

Therefore we have that $\mathcal{C}(G_{\mathtt{JC69}^*}) = (1, 3, 6, 8, 6)$ and $\chi^n(G_{\mathtt{JC69}^*}) = (4^n, 0, 1, 0, 2^n)$. The character table for $G_{\texttt{JC69}^*}$ and χ^n are given by:

$\Omega_{\mathfrak{S}_4}$	id	(AC)(GT)	(ACGT)	(ACG)	(AC)
ω_1	1	1	1	1	1
ω_2	1	1	1	-1	-1
ω_3	2	2	-1	0	0
ω_4	3	-1	0	-1	1
ω_5	3	1	0	1	-1
χ^n	4^n	0	1	0	2^n

.

Kimura 2-parameter model (K80*). The transition matrix for this model was defined in 1.12. The corresponding subgroup is the dihedral group defined in (5.34). Dihedral group has order 8 corresponding to 8 movements that leave a square invariant. Labeling the corners of this square as $\{\texttt{A},\texttt{C},\texttt{G},\texttt{T}\}$, the 8 movements correspond to the following permutations of the corners: rotations and reflections along the horizontal, vertical and diagonal symmetry axes. The group has the following elements:

$$G_{\texttt{K80}^*} = \{id, (\texttt{ACGT}), (\texttt{AG})(\texttt{CT}), (\texttt{ATGC}), (\texttt{AC})(\texttt{GT}), (\texttt{AT})(\texttt{CG}), (\texttt{AG}), (\texttt{CT})\}$$

and is generated by $G_{\texttt{K80}^*} = \langle(\texttt{ACGT}), (\texttt{AG})\rangle$. A set of representatives of the conjugacy classes is: $\{id, (\texttt{AC})(\texttt{GT}), (\texttt{AG})(\texttt{CT}), (\texttt{ACGT}), (\texttt{AG})\}$. Denoting the conjugacy classes in the above order, their elements are given by:

$$\begin{aligned}\mathcal{C}_1 &= \{id\}, \mathcal{C}_2 = \{(\texttt{AC})(\texttt{GT}), (\texttt{AT})(\texttt{GC})\}, \mathcal{C}_3 = \{(\texttt{AG})(\texttt{CT})\},\\ \mathcal{C}_4 &= \{(\texttt{ACGT}), (\texttt{ATGC})\}, \mathcal{C}_5 = \{(\texttt{AG}), (\texttt{CT})\}.\end{aligned}$$

Therefore, we have that $\mathcal{C}(G_{\texttt{K80}^*}) = (1,2,1,2,2)$ and $\chi^n(G_{\texttt{K80}^*}) = (4^n,0,0,0,2^n)$. The character table and the character of the defining representation are given below.

$\Omega_{G_{\texttt{K80}^*}}$	id	(AC)(GT)	(AG)(CT)	(ACGT)	(AG)
ω_1	1	1	1	1	1
ω_2	1	-1	1	1	-1
ω_3	1	-1	1	-1	1
ω_4	1	1	1	-1	-1
ω	2	0	-2	0	0
χ^n	4^n	0	0	0	2^n

Kimura 3-parameter model (K81*). The detailed properties of the model were described in Casanellas and Fernández-Sánchez (2008). The group has 4 elements with 2 generators: $G_{\texttt{K81}^*} = \{id,\ (\texttt{AT})(\texttt{GC}),\ (\texttt{AC})(\texttt{GT}),\ (\texttt{AG})(\texttt{CT})\} = \langle(\texttt{AC})(\texttt{GT}), (\texttt{AG})(\texttt{CT})\rangle$. The generators are given by any pair of distinct and nontrivial group elements. The conjugacy classes are $\{id, (\texttt{AT})(\texttt{CG}), (\texttt{AC})(\texttt{GT}), (\texttt{AG})(\texttt{CT})\}$ and we have that $\mathcal{C}(G_{\texttt{K81}^*}) = (1,1,1,1)$.

Moreover, $\chi^n(G_{\texttt{K81*}}) = (4^n, 0, 0, 0)$ and

$\Omega_{G_{\texttt{K81*}}}$	id	(AT)(CG)	(AC)(GT)	(AG)(CT)
ω_1	1	1	1	1
ω_2	1	-1	-1	1
ω_3	1	-1	1	-1
ω_4	1	1	-1	-1
χ^n	4^n	0	0	0

.

Strand Symmetric Model (SSM). The transition matrix for the strand symmetric model has the form given in 1.15. The associated group has cardinality two and has one generator

$$G_{\texttt{SSM}} = \{id, (\texttt{AT})(\texttt{CG})\} = \langle(\texttt{AT})(\texttt{CG})\rangle.$$

The conjugacy classes for this model are $\{id, (\texttt{AT})(\texttt{CG})\}$, each being a single element, so $\mathcal{C}(G_{\texttt{SSM}}) = (1, 1)$. Lastly, $\chi^n(G_{\texttt{SSM}}) = (4^n, 0)$ and the character table is given by

$\Omega_{G_{\texttt{SSM}}}$	id	(AT)(CG)
ω_1	1	1
ω_2	1	-1
χ^n	4^n	0

General Markov Model The substitution matrices of the GMM model defined in (1.10) do not have any symmetries. Therefore, we associate it with the trivial $G_{\texttt{GMM}} = \langle id\rangle$. There is a single irreducible representation $\omega_1 : G_{\texttt{GMM}} \to \mathbb{C}$ corresponding to the trivial character. The defining representation of $G_{\texttt{GMM}}$ (see Def. 5.56) maps id to the identity linear map in $GL(W)$ so that $\chi^{id}_{\texttt{GMM}} = 4$.

Below we summarize the information for all the models:

- $G = \mathfrak{S}_4$, for the *algebraic Jukes-Cantor model* JC69*,
- $G = \langle(\texttt{ACGT}), (\texttt{AG})\rangle$, for the *algebraic Kimura 2-parameter model* K80*,
- $G = \langle(\texttt{AC})(\texttt{GT}), (\texttt{AG})(\texttt{CT})\rangle$, for the *algebraic Kimura 3-parameter model* K81*,
- $G = \langle(\texttt{AT})(\texttt{CG})\rangle$, for the *strand symmetric model* SSM, and
- $G = \langle id\rangle$, for the *general Markov model* GMM.

Here we prove Lemma 5.23 for the equivariant models. Namely, we will show that the equivariant models are root-independent as dictated by the defintition 5.24. For the SBD this was shown Allman and Rhodes (2006b). This fact for the GMM model can be found e.g. in Allman and Rhodes (2003). We mention already that a choice of the root induces te orientation of the edges. As before, we adopt the convention of the row labels corresponding to the ancestral node of e (more proximal to the root) and the columns to the decendant.

Proof of Lemma 5.23. We present a sketch of the proof for trivalent trees, which can be easily generalized to any phylogenetic tree. We start by assuming a 2-taxon tree, which corresponds to an edge with terminal vertices $e = (u, v)$. If A^e is a Markov matrix and π is stochastic vector, then $D_{\widetilde{\pi}}^{-1}(A^e)^t D_\pi$ are also Markov matrices and $\widetilde{\pi}$ is stochastic. First we check that $(\widetilde{\pi}, \tilde{A}^e)$ is in the model. To see that that $\widetilde{\pi} \in \mathcal{W}_0$ we write

$$P_g\widetilde{\pi} = P_g(A^e)^t\pi = (A^e)^t P_g\pi = (A^e)^t\pi = \widetilde{\pi}.$$

Now, $A^e \in Mod$, so

$$D_{\widetilde{\pi}}^{-1}A^e D_\pi P_g = D_{\widetilde{\pi}}^{-1}A^e P_g D_\pi = D_{\widetilde{\pi}}^{-1}P_g A^e D_\pi = P_g D_{\widetilde{\pi}}^{-1}A^e D_\pi.$$

The equalitites hold because $\pi \in \mathcal{W}_0, A^e \in Mod$ and lastly because $\widetilde{\pi} \in \mathcal{W}_0$. Therefore $D_{\widetilde{\pi}}^{-1}(A^e)^t D_\pi \in Mod.$

Let the root be in u, then $p_{\mathtt{x}_1\mathtt{x}_2} = \pi_{\mathtt{x}_1}A^e_{\mathtt{x}_1,\mathtt{x}_2}$. Upon changing the root to v and defining $\widetilde{\pi}_{\mathtt{x}_2} = \sum_{\mathtt{x}}\pi_{\mathtt{x}}A^e_{\mathtt{x},\mathtt{x}_2}$ and $\tilde{A}^e{}_{\mathtt{x}_2,\mathtt{x}_1} = \widetilde{\pi}^{-1}_{\mathtt{x}_2}A^e_{\mathtt{x}_1,\mathtt{x}_2}\pi_{\mathtt{x}_1}$. Now, we have that $p_{\mathtt{x}_1\mathtt{x}_2} = \widetilde{\pi}_{\mathtt{x}_2}\tilde{A}^e{}_{\mathtt{x}_2,\mathtt{x}_1} = \sum_{\mathtt{x}}\pi_{\mathtt{x}}A^e_{\mathtt{x},\mathtt{x}_2}\widetilde{\pi}^{-1}_{\mathtt{x}_2}A^e_{\mathtt{x}_1,\mathtt{x}_2}\pi_{\mathtt{x}_1}$

In order to see that the result holds for larger trees, we recall the general formula for the parametrization given in (1.1). Now, if we move the root between two adjacent nodes, u and v, the only edge that will change its direction is (u, v), We have shown above that it will not affect the joint probability $\mathbf{p}_{uv}$ and thus, by the formula above, of $\mathbf{p}_{\mathtt{x}_1\ldots\mathtt{x}_n}$. On the other hand, if u and v are not adjacent, there will exist a unique path that joins them. The matrices assigned to the components of this path will be transformed as indicated above. Again, this will not transform the formula for the joint probability. □

Chapter 6

Mixtures and their identifiability

This chapter is a collaboration with Marta Casanellas and Jesús Fernández-Sánchez.

In phylogenetics, it is often assumed that the sites of an alignment are independent and identically distributed. This assumption is not very realistic, however, significantly lowers the number of parameters and makes the inference tractable.

A phylogenetic mixture is a model for which the sites in the alignment belong to a given family or families of distributions. In this section we introduce phylogenetic mixtures from the algebraic point of view. We assume that all sites in the alignment evolve under the same evolutionary model. We prove that the space where distributions from phylogenetic mxitures lie is a linear space. Moreover, we are able to characterize this space for equivariant models via group actions. Lastly, we describe new results on the identifiability of the mixed models in phylogenetics.

6.1 Space of phylogenetic mixtures

Definition 6.1. Fix a set of taxa $[n]$ and an algebraic evolutionary model $\mathcal{M}$. A *phylogenetic mixture (on m-classes)* or *m-mixture* is any vector $p \in \mathcal{L} = \otimes_{[n]} W$ of the form

$$p = \sum_{i=1}^{m} \alpha_i p^i$$

where $p^i \in \mathrm{Im}(\Psi^{\mathcal{M}}_{\mathcal{T}_i})$, $\mathcal{T}_i \in \mathbb{T}_n$ and $\alpha_i \in \mathbb{C}$. As $\Psi^{\mathcal{M}}_{\mathcal{T}_i}$ is a homogeneous map, phylogenetic mixtures are actually vectors of the form $\sum_{i=1}^{m} \check{p}^i$, where $\check{p}^i \in \mathrm{Im}(\Psi^{\mathcal{M}}_{\mathcal{T}_i})$.

Note that on a phylogenetic mixture we allow some (or all) tree topologies $\mathcal{T}_i$ to be the same. In the continuous-time setting we mentioned in Section 1.4, allowing for distinct model parameters on the same topology at different sites is handled by modeling different "classes" of sites by means of the Gamma-rates. In the practical setting, this continuous parameter is discretized and a finite number of classes allowed in the process. Therefore, the discrete Gamma-rates (Γ) with or without the invariable sites are instances of phylogenetic mixtures (we refer to the book Semple and Steel (2003) for an introduction to these concepts).

We denote by $\mathcal{D}_{\mathcal{M}} \subset \mathcal{L}$ the set of all phylogenetic mixtures (on any number of classes) under the algebraic evolutionary model $\mathcal{M}$ and by $\mathcal{D}^m_{\mathcal{M}}$ the set of all phylogenetic mixtures on m-classes.

When we restrict to matrices whose rows sum to one so that we consider the parameterization $\phi_T^{\mathcal{M}}$, one has to restrict the phylogenetic mixtures to points of the form

$$q = \sum_{i=1}^{m} \alpha_i q^i \quad \text{where} \quad q^i \in \mathrm{Im}(\phi_{T_i}^{\mathcal{M}}) \text{ and } \sum_i \alpha_i = 1.$$

We call $\mathcal{D}_{s\mathcal{M}}$ the space of these *stochastic phylogenetic mixtures.*

The following result was proven by Matsen, Mossel and Steel in Matsen et al. (2008) for the two state random cluster model.

Lemma 6.2. Given a set of taxa $[n]$ and an algebraic evolutionary model $\mathcal{M}$, the set of all phylogenetic mixtures $\mathcal{D}_{\mathcal{M}}$ is a vector subspace of $\mathcal{L}$. Similarly, the space $\mathcal{D}_{s\mathcal{M}}$ is a linear variety of the affine space $\mathcal{L}$ contained in the hyperplane H.

Proof. $\mathcal{D}_{\mathcal{M}}$ is a $\mathbb{C}$-vector space by definition.

In order to prove that $\mathcal{D}_{s\mathcal{M}}$ is a linear variety, let q_0 be any point in $\mathcal{D}_{s\mathcal{M}}$, so that $q_0 = \sum_{i=1}^m \alpha_i q^i$ with $q^i \in \mathrm{Im}(\phi_{T_i}^{\mathcal{M}})$, $i = 1, \ldots, m$, and $\sum_i \alpha_i = 1$. Then we can write

$$\mathcal{D}_{s\mathcal{M}} = q_0 + F, \text{ where } F = \{\overrightarrow{q_0 q} \mid q \in \mathcal{D}_{s\mathcal{M}}\}.$$

We only have to show that F is a $\mathbb{C}$-vector space:

1) Let $v = \overrightarrow{q_0 q}$ be a vector in F, then $\lambda v = \overrightarrow{q_0 q'}$ where $q' = q_0 + \lambda \overrightarrow{q_0 q}$. This last point is in $\mathcal{D}_{s\mathcal{M}}$: if $q = \sum_{j=1}^l \beta_j \hat{q}^j$ with $\sum_j \beta_j = 1$, then $q' = (1-\lambda)\sum_{i=1}^m \alpha_i q^i + \lambda \sum_{j=1}^l \beta_j \hat{q}^j$ and the scalar coefficients sum to one $(1-\lambda)\sum_i \alpha_i + \lambda \sum_j \beta_j = (1-\lambda) + \lambda = 1$. Therefore λv is in F.

2) Let $v_1 = \overrightarrow{q^0 q^1}$ and $v_2 = \overrightarrow{q^0 q^2}$ be two vectors in F,

$$q^1 = \sum_j \beta_j \hat{q}_j \quad \text{with} \quad \sum \beta_j = 1,$$
$$q^2 = \sum_k \gamma_k \check{q}_k \quad \text{with} \quad \sum \gamma_k = 1,$$

then $v_1 + v_2 = \overrightarrow{q^0 q'}$ with $q' = \sum_j \beta_j \hat{q}_j + \sum_k \gamma_k \check{q}_k - \sum_i \alpha_i q_i$, and all coefficients together sum to one: $\sum_j \beta_j + \sum_k \gamma_k - \sum_i \alpha_i = 1$.

□

Remark 6.3. By virtue of the previous lemma, $\mathcal{D}_{\mathcal{M}}$ is an algebraic variety that contains $\mathrm{Im}\Psi_T^{\mathcal{M}}$ for any tree T and therefore, it also contains $CV_T^{\mathcal{M}}$. It follows that $\mathcal{D}_{\mathcal{M}}$ equals the set of points of the form $p = \sum p^i$ where $p^i \in CV_{T_i}^{\mathcal{M}}$. Similarly, $\mathcal{D}_{s\mathcal{M}}$ equals the set of points of the form $q = \sum \alpha_i q_i$, where $q_i \in V_{T_i}^{\mathcal{M}}$ and $\sum_i \alpha_i = 1$.

For technical reasons needed in the next result, we introduce the following spaces:

Definition 6.4. Define $\overline{\mathcal{D}_{\mathcal{M}}^m}$ as the set of points p of the form $p = \sum_{i=1}^m p^i$ where $p^i \in CV_{T_i}^{\mathcal{M}}$, and $\overline{\mathcal{D}_{s\mathcal{M}}^m}$ as the set of points q of the form $q = \sum_{i=1}^m \alpha_i q^i$ where $q^i \in V_{T_i}^{\mathcal{M}}$ and $\sum_{i=1}^m \alpha_i = 1$.

Lemma 6.5. *The following equalities hold:*

(a) $\overline{\mathcal{D}^m_{s\mathcal{M}}} = \overline{\mathcal{D}^m_{\mathcal{M}}} \cap H$

(b) $\mathcal{D}_{s\mathcal{M}} = \mathcal{D}_{\mathcal{M}} \cap H$.

Proof. (a) For any $p \in \mathcal{L}$, define $\lambda(p) = \sum_{x_i \in B} p_{x_1,\ldots,x_n}$. Let $q \in \overline{\mathcal{D}^m_{s\mathcal{M}}}$. Then, we can write $q = \sum_{i=1}^m \alpha_i q^i$ for some $q^i \in V^{\mathcal{M}}_{\mathcal{T}_i}$ and $\sum \alpha_i = 1$. Clearly, $q \in \overline{\mathcal{D}^m_{\mathcal{M}}}$. Moreover, $\lambda(q) = \sum_i \alpha_i \lambda(q^i) = \sum_i \alpha_i = 1$. Thus, $q \in H$.

Conversely, let $p = \sum_{i=1}^m p^i$ with $p^i \in CV^{\mathcal{M}}_{\mathcal{T}_i}$ for certain tree topologies $\mathcal{T}_i$, and assume that $\lambda(p) = 1$. Apply Proposition 2.19 of Casanellas et al. (2011) to each p^i to get $p^i = \lambda(p^i) q_i$ for some $q_i \in V^{\mathcal{M}}_{\mathcal{T}_i}$. Then, we have

$$p = \sum_i p^i = \sum_i \lambda(p^i) q_i$$

and $1 = \lambda(p) = \sum_i \lambda(p^i)\lambda(q_i) = \sum_i \lambda(p^i)$ since each q_i lies on H. This proves that $p \in \overline{\mathcal{D}^m_{s\mathcal{M}}}$.

(b) can be proven using (a) and Remark 6.3. □

6.2 The space of phylogenetic mixtures for the equivariant models

This section will be devoted to give a precise description of the space $\mathcal{D}_{\mathcal{M}}$ for the equivariant models $\mathcal{M}$ listed in 5.59. This is precisely the characterization of the space $\mathcal{D}_{\mathcal{M}}$ that will be used in chapter 9 for designing a model selection algorithm. Thus, we will assume that $B = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, $k = 4$ and $W = \langle B \rangle_{\mathbb{C}}$. From now on $\mathcal{L} = \otimes^n W$.

Let $G \leqslant \mathfrak{S}_4$ be a permutation group. We consider the restriction to G of the *defining* representation

$$\rho : \mathfrak{S}_4 \to GL(W) \tag{6.1}$$

given by the permutation of the elements of B. This representation induces a G-module structure on W by setting

$$g \cdot \mathtt{x} := \rho(g)(\mathtt{x}) \in \mathcal{W}.$$

In fact, ρ induces a G-module structure on $\mathcal{L} = \otimes^n \mathcal{W}$ by setting

$$g \cdot (\mathtt{x}_1 \otimes \ldots \otimes \mathtt{x}_n) := g \cdot \mathtt{x}_1 \otimes \ldots \otimes g \cdot \mathtt{x}_n. \tag{6.2}$$

and extends by linearity. According to Notation 5.21, if $\mathtt{X} \in B$ and $g \in G$, $g\mathtt{X}$ will stand for the action of g on $\mathtt{X}$ as introduced above. From now on, the space $\mathcal{L}$ will be implicitly considered as a G-module with this action.

Definition 6.6. Given a set of taxa $[n]$, a *G-tensor* on $[n]$ is an $[n]$-tensor invariant by the action defined in (6.2). The set of G-tensors will be denoted by $\mathcal{L}^G$. If $\mathcal{M}$ is an equivariant evolutionary model, $\mathcal{L}^{G_{\mathcal{M}}}$ will be also denoted by $\mathcal{L}^{\mathcal{M}}$.

The following Theorem describes the set of phylogenetic mixtures for equivariant models.

Theorem 6.7. *If $\mathcal{M}$ is one of the equivariant evolutionary models `JC69`, `K80`, `K81`, `SSM` or `GMM`, then the space of phylogenetic mixtures $\mathcal{D}_{\mathcal{M}}$ coincides with $\mathcal{L}^{G_{\mathcal{M}}}$ and $\mathcal{D}_{s\mathcal{M}} = \mathcal{L}^{G_{\mathcal{M}}} \cap H$.*

This theorem allows one to identify the set of all phylogenetic mixtures $\mathcal{D}_{\mathcal{M}}$ with $\mathcal{L}^{G_{\mathcal{M}}}$, which is a vector subspace of $\mathcal{L}$ whose linear equations are easy to describe, as we will see afterwards in this section. In other words, $\mathcal{L}^{G_{\mathcal{M}}}$ is the space where data coming from any mixture of trees evolving under model $\mathcal{M}$ lies. One can therefore use $\mathcal{L}^{G_{\mathcal{M}}}$ to select the most suitable model for given data. This is the subject of the next chapter 9, where the implementation of the algorithm proposed here is discussed, together with an extensive performance study on simulated and real-life data.

Proof of Theorem 6.7. In Lemma 6.2 we proved that $\mathcal{D}_{\mathcal{M}}$ is a vector subspace of $\mathcal{L}$. Moreover, as we are considering equivariant models, we have $\operatorname{Im}(\Psi_T^{\mathcal{M}}) \subset \mathcal{L}^{G_{\mathcal{M}}}$ for any tree T (see Lemma 4.3 of Draisma and Kuttler (2009)) and hence $\mathcal{D}_{\mathcal{M}}$ is contained in the vector subspace $\mathcal{L}^{G_{\mathcal{M}}}$.

In order to show that $\mathcal{L}^{G_{\mathcal{M}}} = \mathcal{D}_{\mathcal{M}}$ it remains to prove that there does not exist any hyperplane Π containing $\mathcal{D}_{\mathcal{M}}$ and not containing $\mathcal{L}^{G_{\mathcal{M}}}$. If such a hyperplane existed, then it would contain, in particular, all points in $\operatorname{Im}\Psi_{\mathcal{T}}^{\mathcal{M}}$ for any tree topology $\mathcal{T}$. As Π is an algebraic variety, this implies that Π contains $CV_{\mathcal{T}}^{\mathcal{M}}$ for any tree topology $\mathcal{T}$.

It is enough to prove that, for the equivariant models considered here, there are no homogeneous linear polynomials vanishing on all tree topologies, except the linear equations vanishing on $\mathcal{L}^{G_{\mathcal{M}}}$. This result is already known in the literature: for G corresponding to the `GMM` this result was shown in Allman and Rhodes (2008a); for the `SSM` in Casanellas and Sullivant (2005) and for `JC69`*, `K81`*, `K80`* in Sturmfels and Sullivant (2005). In fact, in the case of the `JC69`* and `K80`* models there exist other linear relations, however, they correspond to phylogenetic invariants– these are the equations that vanish on $\Psi_{\mathcal{T}}^{\mathcal{M}}$ for a particular tree topology $\mathcal{T}$ but not for all topologies). The main result in Draisma and Kuttler (2009) comprises all these results.

The equality $\mathcal{D}_{s\mathcal{M}} = \mathcal{L}^{G_{\mathcal{M}}} \cap H$ follows immediately from Lemma 6.5 and the first assertion in this theorem. □

6.3 Equations for the space $\mathcal{L}^{G_{\mathcal{M}}}$

The goal of this section is to compute the dimension of $\mathcal{L}^{G_{\mathcal{M}}}$ where $\mathcal{M}$ is one of the equivariant models listed in Definition 5.59. In addition, we show how a set of independent linear equations defining this space can be obtained. The definitions and facts from group and group representation theory required fo this task can be found in Section 5.4.

Notation 6.8. We recall that $\mathfrak{S}_4$ is a symmetric group on 4 letters (see Definition (5.34)). Denote by *id* the trivial permutation of $\mathfrak{S}_4$. Let $G \leqslant \mathfrak{S}_4$ be a permutation group and $g \in G$. We recall that the *conjugacy class* of g is $C(g) = \{h^{-1}gh : h \in G\}$. If $g_1, g_2 \in G$ (see Def. 5.50). The conjugacy classes $(C_i)_{i=1}^s$ are disjoint and we write $\mathcal{C}(G) = (|C_1|, \dots, |C_s|)$ for the s-tuple of the cardinalities ($\sum_{i=1}^s |C_i| = |G|$). We write χ^n for the character of G associated to the defining representation $G \to GL(\otimes^n \mathcal{W})$ (see Def. 5.48) and represent χ^n by a s-tuple $(t_1, \dots, t_s)$ where $t_i = \chi^n(g)$ for any $g \in C_i$.

Table 6.1 summarizes information about subgroups associated to the equivariant models listed in definition 5.59.

Let $\Omega_G = \{\omega\}_{i=1,\dots,t}$ be a set of the irreducible characters of G, where ω_1 stands for the trivial character. By Maschke's Theorem 5.55 applied to the action of G we write the decomposition of $\otimes^n \mathcal{W}$ into its isotypic components:

$$\otimes^n \mathcal{W} = \oplus_{i=1}^t (\otimes^n W)[\omega_i] \tag{6.3}$$

We recall that each $(\otimes^n \mathcal{W})[\omega_i]$ is isomorphic to a number of copies of the irreducible representation N_i associated to ω_i, $(\otimes^n \mathcal{W})[\omega_i] \cong N_i \otimes \mathbb{C}^{m_i(n)}$ for some positive integer $m_i(n)$, called the *multiplicity of* $\otimes^n W$ *relative to* ω_i. Moreover, by Theorem 5.54, it is known that the set Ω_G forms an orthonormal basis of the space of characters relative to the inner product defined by (5.4).

Using the facts listed above, we can calculate the dimension of the spaces $\dim \mathcal{L}^G$ for the equivariant models. This is summarized in the following Proposition.

Proposition 6.9. *We have*

(i) $\dim \mathcal{L}^{\texttt{SSM}} = 2^{2n-1}$.

(ii) $\dim \mathcal{L}^{\texttt{K81}^*} = 4^{n-1}$

(iii) $\dim \mathcal{L}^{\texttt{K80}^*} = 2^{2n-3} + 2^{n-2}$

(iv) $\dim \mathcal{L}^{\texttt{JC69}^*} = \frac{2^{2n-3}+1}{3} + 2^{n-2}$.

Proof. Let $\mathcal{M}$ be either SSM, K81*, K80* or JC69*. First of all, notice that the space of $G_{\mathcal{M}}$-tensors is just the isotypic component of $\otimes^n \mathcal{W}$ associated to the trivial representation, or equivalently, to the trivial character ω_1:

$$\mathcal{L}^{\mathcal{M}} = (\otimes^n \mathcal{W})[\omega_1].$$

Table 6.1: Group theoretic description of the equivariant evolutionary models ($\texttt{JC69}^*, \texttt{K80}^*, \texttt{K81}^*, \texttt{SSM}$) (see Section 5.5).

$G \leqslant \mathfrak{S}_4$	$\mathcal{M}$	representants of conj. classes	$\mathcal{C}(G)$	$(t_1, \dots, t_s)$
$\langle\texttt{(AT)(CG)}\rangle$	SSM	$\{id, \texttt{(AT)(CG)}\}$	$(1,1)$	$(4^n, 0)$
$\langle\texttt{(AC)(GT)}, \texttt{(AG)(CT)}\rangle$	$\texttt{K81}^*$	$\{id, \texttt{(AT)(CG)}, \texttt{(AC)(GT)}, \texttt{(AG)(CT)}\}$	$(1,1,1,1)$	$(4^n, 0, 0, 0)$
$\langle\texttt{(ACGT)}, \texttt{(AG)}\rangle$	$\texttt{K80}^*$	$\{id, \texttt{(AC)(GT)}, \texttt{(AG)(CT)}, \texttt{(ACGT)}, \texttt{(AG)}\}$	$(1,2,1,2,2)$	$(4^n, 0, 0, 0, 2^n)$
$\mathfrak{S}_4$	$\texttt{JC69}^*$	$\{id, \texttt{(AC)(GT)}, \texttt{(ACGT)}, \texttt{(AG)}, \texttt{(ACG)}\}$	$(1,3,6,8,6)$	$(4^n, 0, 1, 0, 2^n)$

Table 6.2: Orbit composition and their cardinalities given by Lemma 6.14; here ... denotes the set on the left and " repeats the elements of the cell above.

	$\{\texttt{X}\}_{\texttt{GMM}}$	$\{\texttt{X}\}_{\texttt{SSM}}$	$\{\texttt{X}\}_{\texttt{K81}^*}$	$\{\texttt{X}\}_{\texttt{K80}^*}$	$\{\texttt{X}\}_{\texttt{JC69}^*}$
$\mathcal{B}_0$	$\{\texttt{X}\}$	$\dots \cup \{\texttt{(AT)(CG)X}\}$	$\dots \cup \{\texttt{(AC)(GT)X}\}_{\texttt{SSM}}$	...	...
$\mathcal{B}_{\texttt{AG}\mid\texttt{CT}}$	"	"	"	...	$\dots \cup \{\texttt{(AC)X}\}_{\texttt{K80}^*}$
$\mathcal{B}_{\texttt{AC}\mid\texttt{GT}}$	"	"	"	$\dots \cup \{\texttt{(AG)X}\}_{\texttt{K81}^*}$	$\dots \cup \{\texttt{(AT)X}\}_{\texttt{K80}^*}$
$\mathcal{B}_{\texttt{AT}\mid\texttt{CG}}$	"	"	"	$\dots \cup \{\texttt{(AG)X}\}_{\texttt{K81}^*}$	$\dots \cup \{\texttt{(AC)X}\}_{\texttt{K80}^*}$
$\mathcal{B} \setminus \mathcal{B}_2$	"	"	"	$\dots \cup \{\texttt{(AG)X}\}_{\texttt{K81}^*}$	$\dots \cup \{\texttt{(AC)X}\}_{\texttt{K80}^*} \cup \{\texttt{(AT)X}\}_{\texttt{K80}^*}$

Since the dimension of the trivial representation is one, it follows that the dimension of $\mathcal{L}^{\mathcal{M}}$ is precisely the multiplicity $m_1(n)$, that is, the number of times the trivial representation appears in the decomposition of $\otimes^n W$ into isotypic components. As seen in equation (5.6) and thus in (5.7), this multiplicity equals

$$m_1(n) = \langle \chi^n, \omega_1 \rangle = \frac{1}{|G|} \sum_{g \in G} \chi^n(g)\omega_1(g) = \frac{1}{|G|} \sum_{i=0}^{s} |C_i| t_i.$$

The last equality follows from grouping the elements of G in their respective conjugacy classes. Applying the above formula to the subgoups describing the models we obtain the result. For instance

$$\dim \mathcal{L}^{\mathtt{SSM}} = \frac{1}{2} 4^n = 2^{2n-1}.$$

□

Next we provide a set of independent linear equations for $\mathcal{L}^{G_{\mathcal{M}}}$. We will denote a set of patterns of n letters as introduced in 5.21. For notational convenience we further write:

Notation 6.10. We will consider the following subsets of $\mathcal{B} = B^n$:

$$\begin{aligned}
\mathcal{B}_0 &= \{(\mathtt{A}, \ldots, \mathtt{A}), (\mathtt{C}, \ldots, \mathtt{C}), (\mathtt{G}, \ldots, \mathtt{G}), (\mathtt{T}, \ldots, \mathtt{T})\} \\
\mathcal{B}_{\mathtt{AC|GT}} &= \{\mathtt{A}, \mathtt{C}\}^n \cup \{\mathtt{G}, \mathtt{T}\}^n \\
\mathcal{B}_{\mathtt{AG|CT}} &= \{\mathtt{A}, \mathtt{G}\}^n \cup \{\mathtt{C}, \mathtt{T}\}^n \\
\mathcal{B}_{\mathtt{AT|CG}} &= \{\mathtt{A}, \mathtt{T}\}^n \cup \{\mathtt{C}, \mathtt{G}\}^n \\
\mathcal{B}_2 &= \mathcal{B}_{\mathtt{AC|GT}} \cup \mathcal{B}_{\mathtt{AG|CT}} \cup \mathcal{B}_{\mathtt{AT|CG}}.
\end{aligned}$$

The set $\mathcal{B}_0$ is composed of all n-words with only one letter and it is contained in $\mathcal{B}_{\mathtt{AC|GT}}$, $\mathcal{B}_{\mathtt{AG|CT}}$ and $\mathcal{B}_{\mathtt{AT|CG}}$. Similarly, $\mathcal{B}_2$ is composed of all n-words with two letters at most. It is straightforward to check that $|\mathcal{B}_{\mathtt{AC|GT}}| = |\mathcal{B}_{\mathtt{AG|CT}}| = |\mathcal{B}_{\mathtt{AT|CG}}| = 2^{n+1}$ and $|\mathcal{B}_2| = 3 \cdot 2^{n+1} - 8$.

We will adopt multiplicative notation for the n-words in the alphabet B. For instance, we will write $\mathtt{C}^l$ to mean the word and $\underbrace{\mathtt{C} \ldots \mathtt{C}}_{l}$ and $(\mathtt{A}^l)(\mathtt{G}^m)\mathtt{x}_{l+m+1} \ldots \mathtt{x}_n$ to mean $\underbrace{\mathtt{A} \ldots \mathtt{A}}_{l}\underbrace{\mathtt{G} \ldots \mathtt{G}}_{m}\mathtt{x}_{l+m+1} \ldots \mathtt{x}_n$, where $\mathtt{x}_{l+m+1}, \ldots, \mathtt{x}_n$ represent any possible choice of letters.

The main result of this section is the following:

Theorem 6.11. *A set of linearly independent equations $\mathbb{E}^{\mathcal{M}}$ for $\mathcal{L}^{G_{\mathcal{M}}}$ is given by*

$\mathbb{E}^{\mathtt{SSM}}$ *: $p_{\mathtt{X}} = p_{(AT)(CG)\mathtt{X}}$ where $\mathtt{X}$ has $\mathtt{x}_1 \in \{A, C\}$;*

$\mathbb{E}^{\mathtt{K81}^*}$ *: the equations in $\mathbb{E}^{\mathtt{SSM}}$, together with $\boldsymbol{p}_{\mathtt{X}} = \boldsymbol{p}_{(AC)(GT)\mathtt{X}}$, where $\mathtt{X}$ has $\mathtt{x}_1 = A$;*

$\mathbb{E}^{\mathtt{K80}^*}$ *: the equations in $\mathbb{E}^{\mathtt{K81}^*}$, together with $\boldsymbol{p}_{\mathtt{X}} = \boldsymbol{p}_{(AG)\mathtt{X}}$,where $\mathtt{X} \in \mathcal{B} \setminus \mathcal{B}_{AC|GT}$ has $\mathtt{x}_1 = A$, and if T appears in $\mathtt{X}$, there is some C in a preceding position;*

$\mathbb{E}^{\mathtt{JC69}^*}$ *: the equations in $\mathbb{E}^{\mathtt{K80}^*}$, together with $\boldsymbol{p}_{\mathtt{X}} = \boldsymbol{p}_{(AT)\mathtt{X}}$, where $\mathtt{X} \in \mathcal{B}_{AC|GT} \setminus \mathcal{B}_0$ has the form*

$(A^l)(C^m)\mathtt{x}_{l+m+1}\dots\mathtt{x}_n$; and equations $\mathbf{p}_{\mathtt{X}} = \mathbf{p}_{(AC)\mathtt{X}}$ and $p_{\mathtt{X}} = p_{(AT)\mathtt{X}}$ where $\mathtt{X} \in \mathcal{B} \setminus \mathcal{B}_2$ has the form $(A^l)(C^m)\mathtt{x}_{l+m+1}\dots\mathtt{x}_n$ and, if T appears in $\mathtt{X}$, there is some G in a preceding position.

The number of equations added in each case is:

$$\mathtt{SSM} : 2^{2n-1}; \mathtt{K81}^* : 2^{2n-2}; \mathtt{K80}^* : 2^{2n-3} - 2^{n-2}; \mathtt{JC69}^* : 2^{n-1} - 1 + 2\left(\frac{2^{2n-3}+1}{3} - 2^{n-2}\right)$$

.

In order to prove this theorem we refer to a few technical results shown in 5.4. Firstly, by Lemma 5.58 we have that if $G = \langle g_1, \dots, g_t\rangle$, then $\mathcal{L}^G = \bigcap_{i=1}^t \mathcal{L}^{\langle g_i\rangle}$. As a consequence of the above, the system of linear equations for $\mathcal{L}^G$ is obtained from a system of generators of G. That is to say, given a point $\mathbf{p} \in \mathcal{L}$, we have that

$$\mathbf{p} \in \mathcal{L}^G \quad \Leftrightarrow \quad \mathbf{p}_{g\mathtt{X}} = \mathbf{p}_{\mathtt{X}},\ \forall g \in G, \forall\, \mathtt{X} \in \mathcal{B}.$$

Let H be a subgroup of G and $H \setminus G$ the set of right cosets of H in G (see Def. 5.35) We recall that by Lagrange's theorem (5.38): $|H \setminus G| = |G|/|H|$. Moreover, as seen in , the following holds: if $[G : H]$ is the *index* of H in G (Definition 5.37) and $\{g_1, \dots, g_{[G:H]}\}$ is a transversal of $H \setminus G$ (see Def. 5.40), we have a partition of G

$$G = \bigcup_{i=1}^{[G:H]} Hg_i. \tag{6.4}$$

The right cosetcan be understood as a single G-orbit with the natural action of G on it.

Example 6.12. We list the transveral sets for the equivariant models considered here:

1. $[G_{\mathtt{SSM}} : \langle id\rangle] = 2$; a transversal of $\langle id\rangle \setminus G_{\mathtt{SSM}}$ is $\{id, (\mathtt{AT})(\mathtt{CG})\}$.
2. $[G_{\mathtt{K81}^*} : G_{\mathtt{SSM}}] = 2$; a transversal of $G_{\mathtt{SSM}} \setminus G_{\mathtt{K81}^*}$ is $\{id, (\mathtt{AC})(\mathtt{GT})\}$.
3. $[G_{\mathtt{K80}^*} : G_{\mathtt{K81}^*}] = 2$; a transversal of $G_{\mathtt{K81}^*} \setminus G_{\mathtt{K80}^*}$ is $\{id, (\mathtt{AG})\}$.
4. $[G_{\mathtt{JC69}^*} : G_{\mathtt{K80}^*}] = 3$; a transversal of $G_{\mathtt{K80}^*} \setminus G_{\mathtt{JC69}^*}$ is $\{id, (\mathtt{AC}), (\mathtt{AT})\}$.

Notation 6.13. The orbit of $\mathtt{X} \in \mathcal{B}$ under the action of G: $\{\mathtt{X}\}_G = \{g\mathtt{X} : g \in G\}$ is denoted in the literature by $\{\mathtt{X}\}_G$. For clarity of exposition, we will write $\{\mathtt{X}\}_{\mathcal{M}}$, whenever the subgroup G defines the model $\mathcal{M}$.

By Lemma 5.43 we have that if $g_1, \dots, g_m$ is a transversal of $H \setminus G$, then for every $\mathtt{X} \in \mathcal{B}$

$$\{\mathtt{X}\}_G = \bigcup_{i=1,\dots,m} \{g_i\mathtt{X}\}_H.$$

The following Lemma gives a detailed description of the cardinality of the orbits for the equivariant models.

Lemma 6.14. *Let* $\mathtt{X} \in \mathcal{B}$. *Then,*

SSM: $\{\mathtt{X}\}_{\mathtt{SSM}} = \{\mathtt{X}, (AT)(CG)\mathtt{X}\}$ *and there are* 2^{2n-1} *different orbits.*

K81*: $\{\mathtt{X}\}_{\mathtt{K81}^*} = \{\mathtt{X}\}_{\mathtt{SSM}} \cup \{(AC)(GT)\mathtt{X}\}_{\mathtt{SSM}}$ *has cardinality 4 and there are* 2^{2n-2} *different orbits.*

K80*:
- *If* $\mathtt{X} \in \mathcal{B}_{AG|CT}$ *then* $\{\mathtt{X}\}_{\mathtt{K80}^*} = \{\mathtt{X}\}_{\mathtt{K81}^*}$ *has cardinality 4 and there are* 2^{n-1} *different orbits;*
- *if* $\mathtt{X} \in \mathcal{B} \setminus \mathcal{B}_{AG|CT}$, *then* $\{\mathtt{X}\}_{\mathtt{K80}^*} = \{\mathtt{X}\}_{\mathtt{K81}^*} \cup \{(AG)\mathtt{X}\}_{\mathtt{K81}^*}$ *has cardinality 8 and there are* $2^{2n-3} - 2^{n-2}$ *different orbits.*

JC69*:
- *If* $\mathtt{X} \in \mathcal{B}_0$ *then* $\{\mathtt{X}\}_{\mathtt{JC69}^*} = \{\mathtt{X}\}_{\mathtt{K80}^*}$ *has cardinality 4 and there is only one orbit;*
- *if* $\mathtt{X} \in \mathcal{B}_{AC|GT} \setminus \mathcal{B}_0$ *then* $\{\mathtt{X}\}_{\mathtt{JC69}^*} = \{\mathtt{X}\}_{\mathtt{K80}^*} \cup \{(AT)\mathtt{X}\}_{\mathtt{K80}^*}$ *has cardinality 12 and there are* $2^{n-1} - 1$ *different orbits; moreover, the union of such orbits cover the whole* $\mathcal{B}_2 \setminus \mathcal{B}_0$.
- *if* $\mathtt{X} \in \mathcal{B} \setminus \mathcal{B}_2$ *then* $\{\mathtt{X}\}_{\mathtt{JC69}^*} = \{\mathtt{X}\}_{\mathtt{K80}^*} \cup \{(AC)\mathtt{X}\}_{\mathtt{K80}^*} \cup \{(AT)\mathtt{X}\}_{\mathtt{K80}^*}$ *has cardinality 24 and there are* $\frac{1}{3}(2^{2n-3}+1) - 2^{n-2}$ *different orbits.*

The summary of this result is given in the table **??**.

Proof. We will describe the orbits of the elements $\mathtt{X} \in \mathcal{B}$ under the action of their corresponding groups. For the SSM and K81* models this can be done from the definition of the orbits.

SSM: By the defintition of an orbit we obtain that $\{\mathtt{X}\}_{\mathtt{SSM}} = \{\mathtt{X}\} \cup \{(\mathtt{AT})(\mathtt{CG})\mathtt{X}\}$. Since we have that X is a distinct element to (AT)(CG)X for every X, we have that $|\{\mathtt{X}\}_{\mathtt{SSM}}| = 2$.

K81*: Applying Lemma 5.43, we obtain that $\{\mathtt{X}\}_{\mathtt{K81}^*} = \{\mathtt{X}\}_{\mathtt{SSM}} \cup \{(\mathtt{AC})(\mathtt{GT})\mathtt{X}\}_{\mathtt{SSM}}$.

In the above reasoning we in fact used the fact that for the subgroups $G_{\mathtt{SSM}}$ and $G_{\mathtt{K81}^*}$ the stabilizers are trivial (see Definition 5.41 and Theorem 5.44). The idea of the proof for the remaining models is to systematically apply Lemma 5.43.

K80*: Applying Lemma 5.43, we obtain that

$$\{\mathtt{X}\}_{\mathtt{K80}^*} = \{\mathtt{X}\}_{\mathtt{K81}^*} \cup \{(\mathtt{AG})\mathtt{X}\}_{\mathtt{K81}^*}.$$

If $\mathtt{X} \in \mathtt{b}_{\mathtt{AG|CT}}$, then $\{(\mathtt{AG})\mathtt{X}\}_{\mathtt{K81}^*} = \{\mathtt{X}\}_{\mathtt{K81}^*}$ and $\{\mathtt{X}\}_{\mathtt{K80}^*}$ has cardinality 4. The number of such orbits is

$$\frac{|\mathcal{B}_{\mathtt{AG|CT}}|}{4} = 2^{n-1}.$$

If $\mathtt{X} \notin \mathcal{B}_{\mathtt{AG|CT}}$, then $\{(\mathtt{AG})\mathtt{X}\}_{\mathtt{K81}^*} neq \{\mathtt{X}\}_{\mathtt{K81}^*}$, so $\{\mathtt{X}\}_{\mathtt{K80}^*}$ has cardinality 8. The number of such orbits is

$$\frac{|\mathcal{B} \setminus \mathcal{B}_{\mathtt{AG|CT}}|}{8} = 2^{2n-3} - 2^{n-2}.$$

$\mathtt{JC69}^*$: Lemma 5.43 applies to give

$$\{\mathtt{X}\}_{\mathtt{JC69}^*} = \{\mathtt{X}\}_{\mathtt{K80}^*} \cup \{(\mathtt{AC})\mathtt{X}\}_{\mathtt{K80}^*} \cup \{(\mathtt{AT})\mathtt{X}\}_{\mathtt{K80}^*}.$$

(a) If $\mathtt{X} \in \mathcal{B}_0$, then $\{(\mathtt{AC})\mathtt{X}\}_{\mathtt{K80}^*} = \{(\mathtt{AT})\mathtt{X}\}_{\mathtt{K80}^*} = \{\mathtt{X}\}_{\mathtt{K80}^*}$, so $\{\mathtt{X}\}_{\mathtt{JC69}^*}$ has 4 elements. The number of such orbits is

$$|\mathcal{B}_0|/4 = 1.$$

(b) If $\mathtt{X} \in \mathcal{B}_{\mathtt{AC|GT}} \setminus \mathcal{B}_0$, then $(\mathtt{AT})\mathtt{X} \in \mathcal{B}_{\mathtt{AG|CT}}$ and $\{(\mathtt{AC})\mathtt{X}\}_{\mathtt{K80}^*} = \{\mathtt{X}\}_{\mathtt{K80}^*}$ has cardinality 8. Therefore, $\{\mathtt{X}\}_{\mathtt{JC69}^*} = \{(\mathtt{AT})\mathtt{X}\}_{\mathtt{K80}^*} \cup \{\mathtt{X}\}_{\mathtt{K80}^*}$ has cardinality $4+8=12$. The number of such orbits is

$$|\mathcal{B}_{\mathtt{AC|GT}} \setminus \mathcal{B}_0|/4 = 2^{n-1} - 1.$$

Moreover, the number of words involved in such orbits is

$$12(2^{n-1} - 1) = 3 \cdot 2^{n+1} - 12$$

which is the cardinality of $\mathcal{B}_2 \setminus \mathcal{B}_0$.

(c) Finally, if $\mathtt{X} \notin \mathcal{B}_2$, then the three orbits $\{(\mathtt{AC})\mathtt{X}\}_{\mathtt{K80}^*}$, $\{(\mathtt{AT})\mathtt{X}\}_{\mathtt{K80}^*}$ and $\{\mathtt{X}\}_{\mathtt{K80}^*}$ have 8 elements each and are disjoint. Thus, we obtain that

$$\{\mathtt{X}\}_{\mathtt{JC69}^*} = \{\mathtt{X}\}_{\mathtt{K80}^*} \cup \{(\mathtt{AC})\mathtt{X}\}_{\mathtt{K80}^*} \cup \{(\mathtt{AT})\mathtt{X}\}_{\mathtt{K80}^*}$$

has 24 elements. The number of such orbits is

$$\frac{|\mathcal{B} \setminus \mathcal{B}_2|}{24} = \frac{4^n - 3 \cdot 2^{n+1} + 8}{24} = \frac{2^{2n-3}+1}{3} - 2^{n-2}.$$

This proves the claim.

□

Remark 6.15. Notice that given a subgroup G of $\mathfrak{S}_4$, every orbit $o = \{\mathtt{X}_1, \ldots, \mathtt{X}_m\}$ described above provides a G-tensor (a tensor invariant under the action of G) defined by

$$\Sigma(o) = \sum_{i=1}^{m} \mathtt{X}_i.$$

All these tensors are linearly independent, since each orbit involves different vectors of $\mathcal{B}$. It follows that all together they provide a basis for $\mathcal{L}^G$.

Now, we proceed to prove Theorem 6.11.

Proof of Theorem 6.11. In all these cases, the equations are obtained by taking the corresponding transversals given by example 6.12. Assume we have computed a system of equations for the equivariant model associated with some subgroup $H \leqslant G$.

We note that in order to generate G we can restrict to the permutations added to H. The result of Lemma 5.43 states that every new G-orbits result from gluing of certain H-orbits by the action of these added permutations. In practical terms, this means that given the equations for a model $\mathcal{M}_H$, the additional equations required by $\mathcal{M}_G$ are obtained by taking a transversal $\{g_1 = e, \dots, g_{[G:H]}\}$ of $H \setminus G$:

$$\left.\begin{array}{l} \mathbf{p}_{\mathtt{X}} = \mathbf{p}_{g_2\mathtt{X}} \\ \mathbf{p}_{\mathtt{X}} = \mathbf{p}_{g_3\mathtt{X}} \\ \dots \\ \mathbf{p}_{\mathtt{X}} = \mathbf{p}_{g_{[G:H]}\mathtt{X}}. \end{array}\right\} \text{for all } \mathtt{X} \in \mathcal{B}.$$

To avoid repetitions of equations, we have to choose a single element for every G-orbit. Notice that it may happen that for some $\mathtt{X} \in \mathcal{B}$, $\{g_i\mathtt{X}\}_H = \{g_j\mathtt{X}\}_H$ for $i \neq j$. In that case, the equality $\mathbf{p}_{g_j\mathtt{X}} = \mathbf{p}_{g_j\mathtt{X}}$ already holds in the space $\mathcal{L}^H$ and does not provide any restriction. We have to keep into account this possibility in order to obtain a minimal set of equations. That they form a minimal system of equations will follow from their cardinality and the dimension computation of Proposition 6.9.

SSM: As $G_{\mathtt{SSM}}$ is generated by $(\mathtt{AT})(\mathtt{CG})$, a set of equations defining $\mathcal{L}^{\mathtt{SSM}}$ is

$$\{\mathbf{p}_{\mathtt{X}} = \mathbf{p}_{(\mathtt{AT})(\mathtt{CG})\mathtt{X}} : \mathtt{X} \in \mathcal{B}\}.$$

Each SSM-orbit provides a single equation. In order to avoid repetitions of equations, we take $\mathtt{X}$ with $\mathtt{x}_1 \in \{\mathtt{A}, \mathtt{C}\}$. All together, we obtain 2^{2n-1} equations.

K81*: As $\{id, (\mathtt{AC})(\mathtt{GT})\}$ is a transversal of $G_{\mathtt{K81}^*} \setminus G_{\mathtt{SSM}}$,

$$\{\mathbf{p}_{\mathtt{X}} = \mathbf{p}_{(\mathtt{AC})(\mathtt{GT})\mathtt{X}} : \mathtt{X} \in \mathcal{B}\}.$$

As above, each K81*-orbit gives rise to a single equation. To avoid repetitions, we restrict to $\mathtt{X}$ with $\mathtt{X}_1 = \mathtt{A}$. Therefore, we are adding 2^{2n-2} equations.

K80*: we obtain the equations

$$\{\mathbf{p}_{\mathtt{X}} = \mathbf{p}_{(\mathtt{AG})\mathtt{X}} : \mathtt{X} \in \mathcal{B}\}.$$

If $\mathtt{X} \in \mathcal{B}_{\mathtt{AG|CT}}$, we know that $\{\mathtt{X}\}_{\mathtt{K80}^*} = \{\mathtt{X}\}_{\mathtt{K81}^*}$. These orbits do not give rise to new equations. On the other hand, every orbit $\{\mathtt{X}\}_{\mathtt{K80}^*}$ where $\mathtt{X} \notin \mathcal{B}_{\mathtt{AG|CT}}$, provides a single equation. To avoid repetitions, we take $\mathtt{X}$ with $\mathtt{X}_1 = \mathtt{A}$ and if $\mathtt{T}$ appears in $\mathtt{X}$, there is some $\mathtt{C}$ in a preceding position. Since $\mathtt{X} \notin \mathcal{B}_{\mathtt{AG|CT}}$, the existence and unicity of such an element in every $G_{\mathtt{K80}^*}$-orbit is guaranteed. We are adding $2^{2n-3} - 2^{n-2}$ newequations.

JC69*: we add the equations

$$\{\mathbf{p}_{\mathtt{X}} = \mathbf{p}_{(\mathtt{AC})\mathtt{X}} : \mathtt{X} \in \mathcal{B}\} \cup \{\mathbf{p}_{\mathtt{X}} = \mathbf{p}_{(\mathtt{AT})\mathtt{X}} : \mathtt{X} \in \mathcal{B}\}$$

If $\mathtt{X} \in \mathcal{B}_0$, then $\{\mathtt{X}\}_{\mathtt{K80}^*} = \{(\mathtt{AC})\mathtt{X}\}_{\mathtt{K80}^*} = \{(\mathtt{AT})\mathtt{X}\}_{\mathtt{K80}^*}$, so we obtain nothing new in this case.

If $\mathtt{X} \in \mathcal{B}_{\mathtt{AG|CT}} \setminus \mathcal{B}_0$, we add the equations

$$\mathbf{p}_{\mathtt{X}} = \mathbf{p}_{(\mathtt{AC})\mathtt{X}}.$$

To avoid repetitions, we take $\mathtt{X}$ of the form $(\mathtt{A}^l)\,(\mathtt{C}^m)\,\mathtt{x}_{l+m+1}\dots\mathtt{x}_n$, where $l, m \geqslant 1$: we are adding $2^{n-1}-1$ new equations.
By Lemma 6.14, if $\mathtt{X} \in \mathcal{B}_{\mathtt{AC|GT}} \cup \mathcal{B}_{\mathtt{AT|CG}} \setminus \mathcal{B}_0$, then the corresponding $\mathtt{JC69}^*$-orbit contains elements of $\mathcal{B}_{\mathtt{AG|CT}}$: these orbits do not provide new equations.
Finally, if $\mathtt{X} \notin \mathcal{B}_2$, we add the equations

$$\mathbf{p}_{\mathtt{X}} = \mathbf{p}_{(\mathtt{AC})\mathtt{X}} \qquad \mathbf{p}_{\mathtt{X}} = \mathbf{p}_{(\mathtt{AT})\mathtt{X}}.$$

Each orbit provides a couple of equations. To avoid repetitions, we choose $\mathtt{X}$ of the form $(\mathtt{A}^l)\,(\mathtt{C}^m)\,\mathtt{x}_{l+m+1}\dots\mathtt{x}_n$ (where $l, m \geqslant 1$) and such that if $\mathtt{T}$ appears in $\mathtt{X}$, there is some $\mathtt{G}$ in a preceding position. The number of such equations is $\frac{2^{2n-2}+2}{3} - 2^{n-1}$.

□

Example 6.16. As an example, we compute a minimal system of equations for $\mathtt{SSM}$, $\mathtt{K81}^*$, $\mathtt{K80}^*$ and $\mathtt{JC69}^*$ in the case of $n = 3$ leaves.
Equations for $\mathcal{L}^{\mathtt{SSM}}$: $\mathbb{E}^{\mathtt{SSM}}$ is composed of the following equations:

$$\begin{array}{cccc}
\mathbf{p}_{\mathtt{AAA}} = \mathbf{p}_{\mathtt{TTT}} & \mathbf{p}_{\mathtt{AAC}} = \mathbf{p}_{\mathtt{TTG}} & \mathbf{p}_{\mathtt{AAG}} = \mathbf{p}_{\mathtt{TTC}} & \mathbf{p}_{\mathtt{AAT}} = \mathbf{p}_{\mathtt{TTA}} \\
\mathbf{p}_{\mathtt{ACA}} = \mathbf{p}_{\mathtt{TGT}} & \mathbf{p}_{\mathtt{ACC}} = \mathbf{p}_{\mathtt{TGG}} & \mathbf{p}_{\mathtt{ACG}} = \mathbf{p}_{\mathtt{TGC}} & \mathbf{p}_{\mathtt{ACT}} = \mathbf{p}_{\mathtt{TGA}} \\
\mathbf{p}_{\mathtt{AGA}} = \mathbf{p}_{\mathtt{TCT}} & \mathbf{p}_{\mathtt{AGC}} = \mathbf{p}_{\mathtt{TCG}} & \mathbf{p}_{\mathtt{AGG}} = \mathbf{p}_{\mathtt{TCC}} & \mathbf{p}_{\mathtt{AGT}} = \mathbf{p}_{\mathtt{TCA}} \\
\mathbf{p}_{\mathtt{ATA}} = \mathbf{p}_{\mathtt{TAT}} & \mathbf{p}_{\mathtt{ATC}} = \mathbf{p}_{\mathtt{TAG}} & \mathbf{p}_{\mathtt{ATG}} = \mathbf{p}_{\mathtt{TAC}} & \mathbf{p}_{\mathtt{ATT}} = \mathbf{p}_{\mathtt{TAA}} \\
\mathbf{p}_{\mathtt{CAA}} = \mathbf{p}_{\mathtt{GTT}} & \mathbf{p}_{\mathtt{CAC}} = \mathbf{p}_{\mathtt{GTG}} & \mathbf{p}_{\mathtt{CAG}} = \mathbf{p}_{\mathtt{GTC}} & \mathbf{p}_{\mathtt{CAT}} = \mathbf{p}_{\mathtt{GTA}} \\
\mathbf{p}_{\mathtt{CCA}} = \mathbf{p}_{\mathtt{GGT}} & \mathbf{p}_{\mathtt{CCC}} = \mathbf{p}_{\mathtt{GGG}} & \mathbf{p}_{\mathtt{CCG}} = \mathbf{p}_{\mathtt{GGC}} & \mathbf{p}_{\mathtt{CCT}} = \mathbf{p}_{\mathtt{GGA}} \\
\mathbf{p}_{\mathtt{CGA}} = \mathbf{p}_{\mathtt{GCT}} & \mathbf{p}_{\mathtt{CGC}} = \mathbf{p}_{\mathtt{GCG}} & \mathbf{p}_{\mathtt{CGG}} = \mathbf{p}_{\mathtt{GCC}} & \mathbf{p}_{\mathtt{CGT}} = \mathbf{p}_{\mathtt{GCA}} \\
\mathbf{p}_{\mathtt{CTA}} = \mathbf{p}_{\mathtt{GAT}} & \mathbf{p}_{\mathtt{CTC}} = \mathbf{p}_{\mathtt{GAG}} & \mathbf{p}_{\mathtt{CTG}} = \mathbf{p}_{\mathtt{GAC}} & \mathbf{p}_{\mathtt{CTT}} = \mathbf{p}_{\mathtt{GAA}}
\end{array}$$

Equations for $\mathcal{L}^{\mathtt{K81}^*}$: $\mathbb{E}^{\mathtt{K81}^*}$ is composed of $\mathbb{E}^{\mathtt{SSM}}$ together with

$$\begin{array}{cccc}
\mathbf{p}_{\mathtt{AAA}} = \mathbf{p}_{\mathtt{CCC}} & \mathbf{p}_{\mathtt{AAC}} = \mathbf{p}_{\mathtt{CCA}} & \mathbf{p}_{\mathtt{AAG}} = \mathbf{p}_{\mathtt{CCT}} & \mathbf{p}_{\mathtt{AAT}} = \mathbf{p}_{\mathtt{CCG}} \\
\mathbf{p}_{\mathtt{ACA}} = \mathbf{p}_{\mathtt{CAC}} & \mathbf{p}_{\mathtt{ACC}} = \mathbf{p}_{\mathtt{CAA}} & \mathbf{p}_{\mathtt{ACG}} = \mathbf{p}_{\mathtt{CAT}} & \mathbf{p}_{\mathtt{ACT}} = \mathbf{p}_{\mathtt{CAG}} \\
\mathbf{p}_{\mathtt{AGA}} = \mathbf{p}_{\mathtt{CTC}} & \mathbf{p}_{\mathtt{AGC}} = \mathbf{p}_{\mathtt{CTA}} & \mathbf{p}_{\mathtt{AGG}} = \mathbf{p}_{\mathtt{CTT}} & \mathbf{p}_{\mathtt{AGT}} = \mathbf{p}_{\mathtt{CTG}} \\
\mathbf{p}_{\mathtt{ATA}} = \mathbf{p}_{\mathtt{CGC}} & \mathbf{p}_{\mathtt{ATC}} = \mathbf{p}_{\mathtt{CGA}} & \mathbf{p}_{\mathtt{ATG}} = \mathbf{p}_{\mathtt{CGT}} & \mathbf{p}_{\mathtt{ATT}} = \mathbf{p}_{\mathtt{CGG}}
\end{array}$$

Equations for $\mathcal{L}^{\mathtt{K80}^*}$: $\mathbb{E}^{\mathtt{K80}^*}$ is composed of $\mathbb{E}^{\mathtt{K81}^*}$ together with

$$\begin{array}{ccc}
\mathbf{p}_{\mathtt{AAG}} = \mathbf{p}_{\mathtt{GAA}} & \mathbf{p}_{\mathtt{ACG}} = \mathbf{p}_{\mathtt{GCA}} & \mathbf{p}_{\mathtt{ACT}} = \mathbf{p}_{\mathtt{GCT}} \\
\mathbf{p}_{\mathtt{AGA}} = \mathbf{p}_{\mathtt{GAG}} & \mathbf{p}_{\mathtt{AGC}} = \mathbf{p}_{\mathtt{GAC}} & \mathbf{p}_{\mathtt{AGG}} = \mathbf{p}_{\mathtt{GAA}}
\end{array}$$

Equations for $\mathcal{L}^{\mathtt{JC69}^*}$*:* $\mathbb{E}^{\mathtt{JC69}^*}$ is composed of $\mathbb{E}^{\mathtt{K80}^*}$ together with

$$\mathbf{p}_{\mathtt{AAC}} = \mathbf{p}_{\mathtt{TTC}} \qquad \mathbf{p}_{\mathtt{ACA}} = \mathbf{p}_{\mathtt{TCT}} \qquad \mathbf{p}_{\mathtt{ACC}} = \mathbf{p}_{\mathtt{TCC}} \qquad \mathbf{p}_{\mathtt{ACG}} = \mathbf{p}_{\mathtt{CAG}} \qquad \mathbf{p}_{\mathtt{ACG}} = \mathbf{p}_{\mathtt{TCG}}.$$

To summarize, the results of this section show that the set of probability distributions for the bases at the leaves that come from a mixture of trees under a discrete-time evolutionary model coincides with the set of distributions satisfying a certain collection of linear invariants. Adopting the definition that the mixtures on the same tree topology contain distributions coming from models employing discrete gamma rates (Γ) from Yang (1994) and/or invariable sites (I) (Steel et al., 2000) this is a powerful results with possible applications in model selection. We described an effective algorithm to obtain the linear invariants characterizing phylogenetic mixtures. Chapter 9 presents the implementation of the method and the results on its performance.

6.4 Identifiability of phylogenetic mixtures

Identifiability lies at the core of applicability of any model in virtually any setting involving data analysis or inference. Some of the most comprehensive references for the identifiability problems in phylogenetic models and their mixtures include Chang (1994), Stefankovic and Vigoda (2007) and Allman et al. (2010).

Definition 6.17. Given two projective varieties $X, Y \subset \mathbb{P}^m$, the *join* of X and Y, $X \vee Y$, is the smallest variety in $\mathbb{P}^m$ containing all lines $\overline{xy}$ with $x \in X$, $y \in Y$ and $x \neq y$ (see (Harris, 1992, 8.1) for the details of this definition). Similarly, we can define the join of projective varieties $X_1, \dots, X_h \subset \mathbb{P}^m$, $\vee_{i=1}^h X_i$, as the smallest subvariety in $\mathbb{P}^m$ containing all the linear varieties spanned by $x_1, \dots, x_h$ with $x_i \in X_i$ and $x_i \neq x_j$. It is known that

$$\dim(\vee_{i=1}^h X_i) \leqslant \min\{\sum_{i=1}^h \dim(X_i) + h - 1, m\}.$$

The right hand side of this inequality is usually known as the expected dimension of $\vee_{i=1}^h X_i$.

For example, if we consider the join $\vee_{i=1}^h \mathbb{P}V_{\mathcal{T}_i}^{\mathcal{M}}$ for certain tree topologies $\mathcal{T}_i$ on the leaf set $[n]$ and a given evolutionary model $\mathcal{M}$, then there is a dominant rational map

$$\mathbb{P}V_{\mathcal{T}_1}^{\mathcal{M}} \times \mathbb{P}V_{\mathcal{T}_2}^{\mathcal{M}} \times \ldots \times \mathbb{P}V_{\mathcal{T}_h}^{\mathcal{M}} \times \mathbb{P}^{h-1} \dashrightarrow \vee_{i=1}^h \mathbb{P}V_{\mathcal{T}_i}^{\mathcal{M}} \subset \mathbb{P}(\mathcal{L}).$$

corresponding to the projective closure of the parameterization $\phi_{\mathcal{T}_1} \vee \ldots \vee \phi_{\mathcal{T}_h}$ defined by

$$\begin{array}{rcl} Par_{s\mathcal{M}}(\mathcal{T}_1) \times \ldots \times Par_{s\mathcal{M}}(\mathcal{T}_h) \times \Omega & \longrightarrow & \mathcal{L} \\ ((\xi_1, \dots, \xi_h), \mathbf{a}) & \mapsto & \sum_j a_i \phi_{\mathcal{T}_i}^{\mathcal{M}}(\xi_i) \end{array}$$

where $\Omega = \{\mathbf{a} = (a_1, \dots, a_h) \mid \sum_i a_i = 1\}$ is isomorphic to an affine open subset of $\mathbb{P}^{h-1}$.

In this setting, an h-mixture on $\{\mathcal{T}_1, \dots, \mathcal{T}_h\}$ corresponds to a point in the variety

$\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}^{\mathcal{M}}$. We will use this algebraic variety to study the identifiability of phylogenetic mixtures.

We recall the definition of generic identifiability of the tree topologies on h-mixtures (see for example Allman et al. (2010)).

Definition 6.18. The *tree topologies* on h-mixtures over $\mathcal{M}$ are *generically identifiable* if for any set of trivalent tree topologies $\mathcal{T}_1 \ldots, \mathcal{T}_h$ and generic choice of $(\xi_1, \ldots \xi_h, \mathbf{a}) \in Par_{s\mathcal{M}}(\mathcal{T}_1) \times \ldots \times Par_{s\mathcal{M}}(\mathcal{T}_h) \times \Omega$, the equality

$$\phi_{\mathcal{T}_1} \vee \ldots \vee \phi_{\mathcal{T}_h}(\xi_1, \ldots \xi_h, \mathbf{a}) = \phi_{\mathcal{T}_1'} \vee \ldots \vee \phi_{\mathcal{T}_h'}(\xi_1', \ldots \xi_h', \mathbf{a}'),$$

for tree topologies $\{\mathcal{T}_1', \ldots, \mathcal{T}_h'\}$ and stochastic parameters $(\xi_1', \ldots \xi_h', \mathbf{a}')$, implies

$$\{\mathcal{T}_1 \ldots, \mathcal{T}_h\} = \{\mathcal{T}_1' \ldots, \mathcal{T}_h'\}.$$

In terms of algebraic varieties this is equivalent to saying that the variety $\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}^{\mathcal{M}}$ is not contained in $\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i'}^{\mathcal{M}}$ and vice versa.

The tree topologies are the discrete parameters of h-mixtures. When we come to the continuous parameters we have the following definition.

Definition 6.19. The *continuous parameters* on h-mixtures on $\mathcal{T}_1, \ldots, \mathcal{T}_h$ under an evolutionary model $\mathcal{M}$ are *generically identifiable* if for generic choices of stochastic parameters $(\xi_1, \ldots, \xi_h, \mathbf{a})$, the equality

$$\phi_{\mathcal{T}_1} \vee \ldots \vee \phi_{\mathcal{T}_h}(\xi_1, \ldots \xi_h, \mathbf{a}) = \phi_{\mathcal{T}_1} \vee \ldots \vee \phi_{\mathcal{T}_h}(\xi_1', \ldots \xi_h', \mathbf{a}')$$

for stochastic parameters $(\xi_1', \ldots, \xi_h', \mathbf{a}')$ implies $(\xi_1, \ldots \xi_h, \mathbf{a}) = (\xi_1', \ldots \xi_h', \mathbf{a}')$ or an allowed permutation of the parameters (Allman et al., 2010, Definition 2).

In terms of algebraic varieties, generic identifiability of continuous parameters implies that the generic fibers of the map $\phi_{\mathcal{T}_1} \vee \ldots \vee \phi_{\mathcal{T}_h}$ are finite. In particular, the fiber dimension theorem applies (cf. (Harris, 1992, Thm 11.12)) to obtain

$$\dim\left(\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}\right) = \sum_{i=1}^{h}\dim\left(\mathbb{P}V_{\mathcal{T}_i}\right) + h - 1$$

The converse of this result (that is, finite generic fibers of $\phi_{\mathcal{T}_1} \vee \cdots \vee \phi_{\mathcal{T}_h}$ imply generic identifiability) is not necessarily true because a finite fiber can be formed by more than one point stochastically meaningful.

Example 6.20. The tree topologies and the continuous parameters are generically identifiable for the unmixed equivariant models `JC69`*, `K80`*, `K81`*, `SSM`, `GMM` (see Corollary 3.9 Casanellas and Fernández-Sánchez, 2011).

If the continuous parameters are generically identifiable under an evolutionary model $\mathcal{M}$, then the dimension of the variety $\mathbb{P}V_{\mathcal{T}}^{\mathcal{M}}$ is the same for all trivalent tree topologies on n taxa and corresponds to the number of free parameters of the stochastic model (fiber dimension theorem cf. (Harris, 1992, Theorem 11.12)). Let $d_{\mathcal{M}}$ be this dimension, then we have the following result.

Theorem 6.21. *Let $\mathcal{M}$ be an evolutionary model for which continuous parameters are generically identifiable on trivalent trees and let $h_0 := \frac{\dim \mathcal{D}_{\mathcal{M}}}{d_{\mathcal{M}}+1}$ where $d_{\mathcal{M}}$ is the dimension of $\mathbb{P}V_T^{\mathcal{M}}$ as above. Then either the continuous parameters or the tree parameters are not generically identifiable for h-mixtures under the model $\mathcal{M}$ if $h \geqslant h_0$.*

Remark 6.22. This theorem proves that it makes no sense to do phylogenetic inference for h-mixtures when $h \geqslant h_0$.

Corollary 6.23. *Let $[n]$ be a set of taxa and $\mathcal{M}$ be one of the equivariant models $\texttt{JC69}^*$, $\texttt{K80}^*$, $\texttt{K81}^*$, $\texttt{SSM}$ and $\texttt{GMM}$. Then phylogenetic h-mixtures under these models are not identifiable for $h \geqslant h_0$ where*

- $h_0 = \frac{4^n}{12(2n-3)+4}$ *if* $\mathcal{M} = \texttt{GMM}$,
- $h_0 = \frac{2^{2n-1}}{6(2n-3)+2}$ *if* $\mathcal{M} = \texttt{SSM}$,
- $h_0 = \frac{4^{n-1}}{3(2n-3)+1}$ *if* $\mathcal{M} = \texttt{K81}^*$,
- $h_0 = \frac{2^{2n-3}+2^{n-2}}{2(2n-3)+1}$ *if* $\mathcal{M} = \texttt{K80}^*$,
- $h_0 = \frac{2^{2n-3}+3\cdot 2^{n-2}+1}{3(2n-2)}$ *if* $\mathcal{M} = \texttt{JC69}^*$.

Proof. Theorem 6.7 shows that $\mathcal{L}^{\mathcal{M}} = \mathcal{D}_{\mathcal{M}}$ and Proposition 6.9 gives the dimension of this space in each case. Then, we apply Theorem 6.21 taking into account that $d_{\texttt{GMM}} = 12(2n-3)+3$, $d_{\texttt{SSM}} = 6(2n-3)+1$, $d_{\texttt{K81}^*} = 3(2n-3)$, $d_{\texttt{K80}^*} = 2(2n-3)$ and $d_{\texttt{JC69}^*} = 2n-3$. □

Example 6.24. Consider the Kimura 3-parameter model $\texttt{K81}^*$ and consider trees on $n = 4$ taxa. Then for any $h \geqslant 4$, phylogenetic h-mixtures are not identifiable (Corollary 6.23). We are not aware of any result proving that mixtures of 2 or 3 different tree topologies under this model are identifiable (either for tree parameters or for continuous parameters).

Example 6.25. If we consider the Jukes-Cantor model $\texttt{JC69}^*$ on $n = 4$ taxa, then Corollary 6.23 tells us that for $h \geqslant 3$, h-mixtures are not identifiable. Therefore for this particular model on four taxa the identifiability is solved: the tree and continuous parameters are generically identifiable for the unmixed model; the tree parameters are generically identifiable for 2-mixtures (Allman et al., 2010, Theorem 10); the continuous parameters are generically identifiable for 2-mixtures on different tree topologies and not identifiable for the same tree topology (Allman et al., 2010, Theorem 23); either the continuous parameters or the tree topologies are not generically identifiable for more than two mixtures (Corollary 6.23).

Proof of Theorem 6.21. Let $edim(h) := hd_{\mathcal{M}} + h - 1$. Then the variety $\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}$ has dimension $\leqslant edim(h)$. Indeed, as $\vee_i\phi_{\mathcal{T}_i}$ is a parameterization of an open subset of $\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}$, then the dimension of $\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}$ is less or equal than $\sum \dim \mathbb{P}V_{\mathcal{T}_i} + h - 1$. Moreover, the dimension of $\mathbb{P}V_{\mathcal{T}_i}$ is equal to $d_{\mathcal{M}}$ if $\mathcal{T}_i$ is trivalent (because the continuous parameters for the unmixed models we are considering are generically identifiable) and is less than $d_{\mathcal{M}}$ for non-trivalent trees. Therefore $\dim(\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}) \leqslant edim(h)$.

If we consider only trivalent trees $\mathcal{T}_i$, then $\sum \dim \mathbb{P}V_{\mathcal{T}_i} + h - 1 = edim(h)$ and therefore $\dim(\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}) < edim(h)$ if and only if $\dim(\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}) < \sum \dim \mathbb{P}V_{\mathcal{T}_i} + h - 1$. Moreover, by fiber dimension theorem applied to $\vee\phi_{\mathcal{T}_i}$, equality holds if and only if the generic fiber of $\vee\phi_{\mathcal{T}_i}$ has dimension 0. In particular, if $\dim(\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}) < edim(h)$ then the continuous parameters of this phylogenetic mixture are not identifiable.

If $h_0 = \frac{\dim \mathcal{D}_{\mathcal{M}}}{d_{\mathcal{M}}+1}$ then, $edim(h_0) = h_0(d_{\mathcal{M}} + 1) - 1 = \dim \mathcal{D}_{\mathcal{M}} - 1$. Now we fix an $h \in \mathbb{N}$ with $h \geqslant h_0$, so that one has $edim(h) \geqslant \dim(\mathcal{D}_{\mathcal{M}}) - 1$.

Two things could happen:

(a) For all tree topologies $\{\mathcal{T}_1, \ldots, \mathcal{T}_h\}$ one has $\dim(\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}) < \dim(\mathcal{D}_{\mathcal{M}}) - 1$.

(b) There exists a set of tree topologies $\{\mathcal{T}_1, \ldots, \mathcal{T}_h\}$ for which $\dim(\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}) = \dim(\mathcal{D}_{\mathcal{M}}) - 1$.

Case (a) implies that for any set of trivalent tree topologies $\{\mathcal{T}_1, \ldots, \mathcal{T}_h\}$ one has $\dim(\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}) < edim(h)$. And we have seen above that this implies that the continuous parameters are not generically identifiable.

In case (b) one has that $\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i} = \mathbb{P}(\mathcal{D}_{\mathcal{M}})$. Indeed, $\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i} \subset \mathbb{P}(\mathcal{D}_{\mathcal{M}})$ and $\dim(\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}) = \dim(\mathcal{D}_{\mathcal{M}}) - 1 = \dim(\mathbb{P}(\mathcal{D}_{\mathcal{M}}))$ which implies that both varieties coincide (the proper subvarieties of an affine space have dimension strictly smaller than it). In particular any other h-mixture (which is a point in $\mathbb{P}(\mathcal{D}_{\mathcal{M}})$) would be contained in $\vee_{i=1}^{h}\mathbb{P}V_{\mathcal{T}_i}$ and therefore the topologies are not generically identifiable. □

Remark 6.26. The negative result of Theorem 6.21 should be complemented with the following positive result of Rhodes and Sullivant (2011): if $\mathcal{M} = \texttt{GMM}$ and one restricts to h-mixtures on the same trivalent tree topology $\mathcal{T}$, then the tree topology and the continuous parameters are generically identifiable if $h < 4^{\lceil \frac{n}{4} \rceil - 1}$.

Part III

Implementations and applications

Chapter 7

GenNon-h

In chapter 3 (Casanellas and Kedzierska, 2011) we give algorithms to generate stochastic transition matrices under the equivariant models considered in the thesis, $\mathcal{M} \in \{\texttt{JC69}^*, \texttt{K81}^*, \texttt{K80}^*, \texttt{SSM}, \texttt{GMM}\}$ and given branch lengths of the underlying tree, $\mathcal{T}$. In all models but the GMM, these algorithms provide the full set of stochastic matrices of a given form and branch length (1.2). As shown in Allman and Rhodes (2003), the substitution parameters for the GMM model (and thus for all its submodels), are identifiable up to permutation. This is a source of possible problems in parameter recovery and branch length calculations when the determinant of the substitution matrix can be negative (see e.g. Zou et al. (2011)). We therefore implemented an extended version of the algorithms given in chapter 3, in that we fabricate matrices of the *Diagonal Largest in Column* (*DLC*) type (Chang, 1996). *DLC* matrices have the property that the largest entry in every column is placed on the diagonal. These substitution matrices are close to the so-called "biologically meaningful" substitution matrices in which the diagonal entries are larger than the off-diagonal ones. In addition, they also share an important feature of being identifiable– there exists a unique set of substitution matrices satisfying the *DLC* condition and a unique root distribution that leads to a given joint distribution at the leaves. In other words, data generated under the *DLC* matrices and sufficient alignment sizes have high chances of being identifiable and therefore can be safely used to test hypotheses about the tree or the data.

Thus, the algorithm proceeds as follows: Firstly, given a model $\mathcal{M}$ and a tree $\mathcal{T}$ with assigned branch lengths, for each edge e in $\mathcal{T}$ we generate a matrix of the type $\mathcal{M}$ corresponding to the length of edge e. If the resulting matrix is not *DLC*, we apply a permutation of rows to convert it into a DLC matrix. Every model has a set of permutations allowable such that the structure of the matrix is mainatined. If neither of the permutations creates a DLC matrix, we generate a new matrix and repeat the procedure. Next, given the length of the multiple sequence alignment, we use the matrices fabricated in the previous step to generate a multiple sequence alignment evolving according to the Markov process on $\mathcal{T}$.

The algorithm was implemented as a C++ package called GenNon-h available at http://genome.crg.es/cgi-bin/phylo_mod_sel/AlgGenNonH.pl. GenNon-h takes as an input a tree in a Newick format (rooted or unrooted, nodes can have any order) with annotated branch lengths measured as the expected number of substitutions per site. Other arguments include base name for the output files, length of the alignment and a model. An input line is therefore as follows:

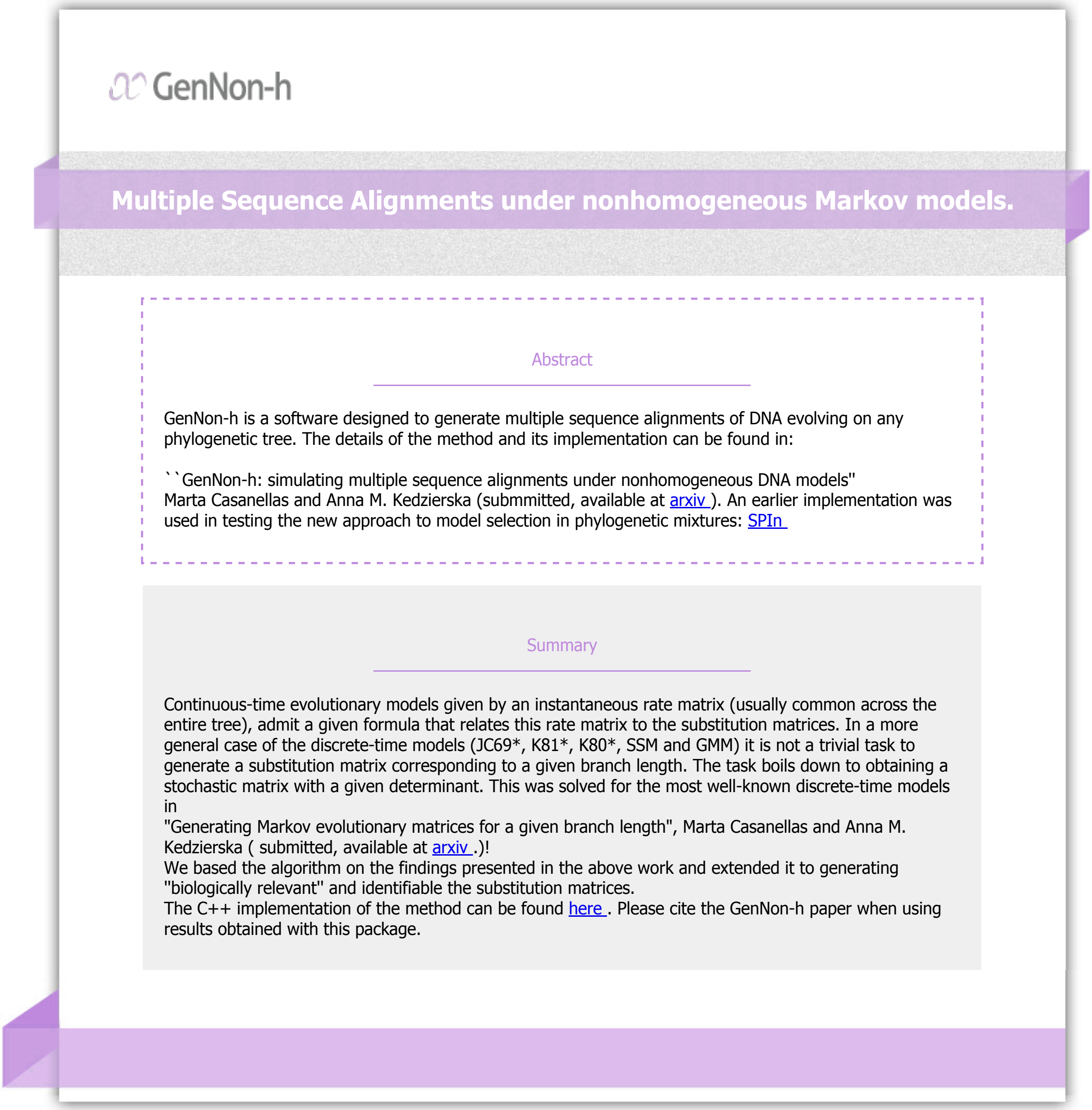
$$\text{GenNon-h } \langle treefile\rangle\langle outputfile\rangle\langle length\rangle\langle model\rangle$$

The output files include a fasta file with the simulated multiple sequence alignment on $\mathcal{T}$ saved under the name specified with an extension ".dat". The file lists the parameters used for the simulations. The order of the matrices corresponds to the order of reading the branches of the Newick format– terminal branches are followed by the edges starting at the root, proceeding from left to right top down (package contains a README file with the detailed information).

Table 7.1: GenNon-h :time to generate 100 alignments of length 1,000bp on a 5-taxon tree on a Macintosh 2.4 GHz Intel Core 2 Duo with 4GB

Model	JC69*	K80*	K81*	SSM	GMM
Time	2.6s	0m2.6s	0m2.5s	2.6s	3.0s

In order to test the speed of GenNon-h, we checked the times it took to generate 100 alignments of $1,000$nt on a tree $((seq1 : 0.01, seq2 : 0.2, seq3 : 0.3) : 0.5, seq4 : 0.4, seq5 : 0.7)$. The results are given in Table 8.2.



GenNon-h

Multiple Sequence Alignments under nonhomogeneous Markov models.

Abstract

GenNon-h is a software designed to generate multiple sequence alignments of DNA evolving on any phylogenetic tree. The details of the method and its implementation can be found in:

``GenNon-h: simulating multiple sequence alignments under nonhomogeneous DNA models"
Marta Casanellas and Anna M. Kedzierska (submmitted, available at arxiv). An earlier implementation was used in testing the new approach to model selection in phylogenetic mixtures: SPIn

Summary

Continuous-time evolutionary models given by an instantaneous rate matrix (usually common across the entire tree), admit a given formula that relates this rate matrix to the substitution matrices. In a more general case of the discrete-time models (JC69*, K81*, K80*, SSM and GMM) it is not a trivial task to generate a substitution matrix corresponding to a given branch length. The task boils down to obtaining a stochastic matrix with a given determinant. This was solved for the most well-known discrete-time models in
"Generating Markov evolutionary matrices for a given branch length", Marta Casanellas and Anna M. Kedzierska (submitted, available at arxiv .)!
We based the algorithm on the findings presented in the above work and extended it to generating "biologically relevant" and identifiable the substitution matrices.
The C++ implementation of the method can be found here . Please cite the GenNon-h paper when using results obtained with this package.

Universitat Politecnica de Catalunya
Center for Genomic regulation
Support and feedback: contact Ania
This is a free software and it can be redistributed, modified or else as given by the terms of the GNU General Public License. Thank you and enjoy!

Chapter 8

Empar

In this section we present the details of the implementation and the performance of `Empar`. `Empar` is an implementation of the Expectation-Maximization algorithm for parameter inference in discrete-time models introduced in chapter 4. The general version of the algorithm is applicable whenever an explicit formula for the MLE can be derived. In the first version of `Empar` we included the equivariant models considered in this thesis (i.e. `JC69`*, `K80`*, `K81`*, `SSM`, see sections 1.4 and 5.2).

8.1 Statistical testing

As the substitution matrices are stochastic with row sums equal to 1, not all of its entries are free to vary. The number d of free parameters for transition matrices in `JC69`*, `K80`*, `K81`*, `SSM` and `GMM` models is 1, 2, 3, 6, and 12 respectively. There are two free parameters for the root distribution on the `SSM` models and three on the `GMM`, whereas the root distribution is uniform for the other models. For clarity in exposition, we omit any reference to root distribution from now on as it can be easily added to the formulae below.

For convenience we adopt the notation of taking off-diagonal entries as free parameters of the model and collecting them into a vector ξ. That is, given a substitution matrix A^e in one of the models above we call $\xi_1 = A_{1,2}$, ξ_2 the next (from left to right and top to bottom) off-diagonal entry that is different from ξ_1 and we keep going until ξ_d. In what follows, ξ_k^e will mean the kth free parameter in matrix A^e associated to edge e.

Let $\xi = (\xi_i^e)_{i=1,\dots,d,e\in E(\mathcal{T})}$ denote a vector of free parameters for an evolutionary model $\mathcal{M}$ as above and let $\hat{\xi}$ be its maximum likelihood estimators (MLEs). The whole set of parameters $\theta = \{\pi, (A^e)_{e\in E(\mathcal{T})}\}$ can be written as a function of the free parameters ξ and we write $\mathcal{L}_{obs}(\xi; u_D) = \prod p_{\mathbf{x}}(\theta(\xi))^{u_{\mathbf{x}}}$ for $\mathcal{L}_{obs}(\theta(\xi); u_D)$, see notation in chapter 4.

Under certain regularity conditions (Zacks (1971)[Chap. 5]) the MLE $\hat{\xi}$ exists, is consistent, efficient and asymptotically normal with mean ξ and the covariance matrix given by the inverse of the Fisher information matrix (the negative of the Hessian matrix) Rao (1973); Efron and Hinkley (1978)[Chap. 6, p. 127]. The entries of the $d|E(\mathcal{T})| \times d|E(\mathcal{T})|$ Fisher information matrix over free parameters are given by:

$$\mathbf{I}(\xi_k^e, \xi_{k'}^{e'}) = -\mathbf{E}\left(\frac{\partial^2 \log L_{obs}(\xi; u_D)}{\partial \xi_k^e \partial \xi_{k'}^{e'}}\right). \tag{8.1}$$

The derivation of the formulae for the Fisher information matrix under the discrete-time models with linear restrictions on the parameters is given in the appendix B.

The Wald statistics for testing the null hypothesis $\xi_i^e = \hat{\xi}_i^e$, $e \in E(\mathcal{T}), i = 1, \dots, d$, is

$$(\hat{\xi^e} - \xi^e)^T \mathbf{I}^e (\hat{\xi^e} - \xi^e) \sim \chi_d^2, \tag{8.2}$$

where $\mathbf{I}^e$ denotes the $d \times d$ slice of $\mathbf{I}$ corresponding to the parameters of $e \in E(\mathcal{T})$. The $p - value$ can thus be easily calculated by looking at the tails of the corresponding χ^2 distribution. Figure 8.1 depicts example of the fitness of the data to the theoretical distribution (8.2).

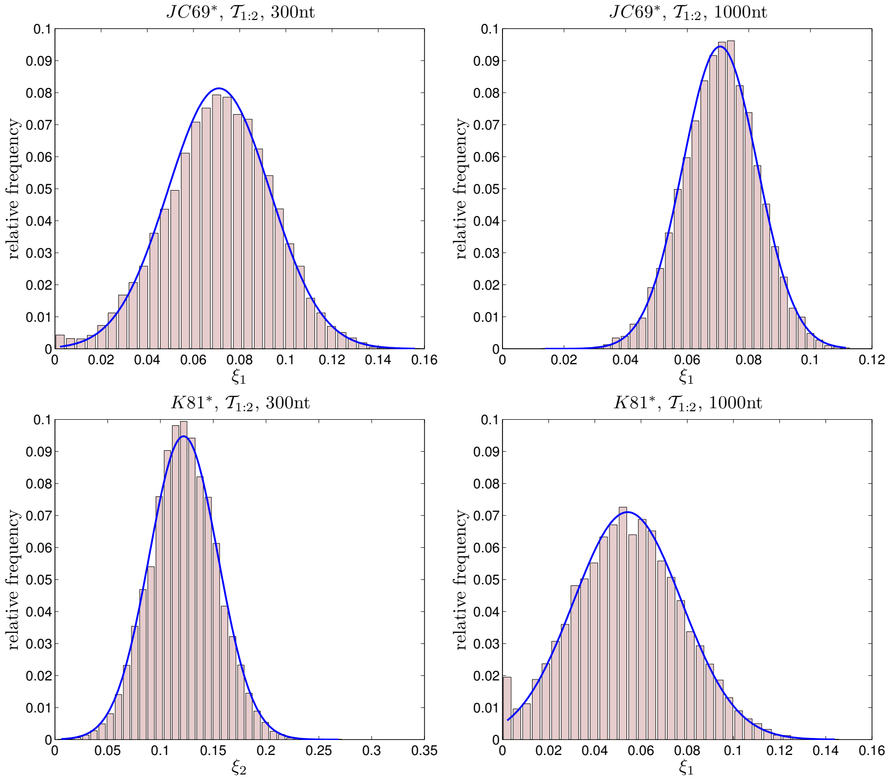


Figure 8.1: Fit of the asymptotic theoretical distribution of the maximum likelihood estimator: examples on the free parameters ξ_1 and ξ_2 of the inner branch in the $\mathcal{T}^4_{\text{balanced}}$ with $l = 0.5$.

In the above, we used the inverse of the Fisher information as an estimate of the covariance matrix for the initial parameters in the test runs. Equivalent derivations hold if we use the observed Fisher information to estimate the covariance for the input data. Variances of the free parameters of the model are provided in the output of `Empar` and the full (observed) covariance matrix is written to an output file. As above, these can be used as plug-in estimators in (8.2) for the calculation of $p - values$ or normal confidence intervals for the parameters.

Theoretical parameter variance

We denote by $V_{i,i}^e$ the i^{th} diagonal entry of the matrix $(\mathbf{I}^e)^{-1}$ corresponding to the variance of the free parameter ξ_i^e, $i = 1, \dots, d$. For the models with $d > 1$ (i.e. all but JC69*), for each edge e we summarized the variances of the free parameters in a combined form cV^e:

$$cV^e(\xi^e) = \frac{\sum_{j=1}^{d} \left(V^e_{j,j} + \left(\xi^e_j - \frac{\sum_{j=1}^{d} \xi^e_j}{d} \right)^2 \right)}{d}. \tag{8.3}$$

8.2 Synthetic data

In order to assess the accuracy of the method proposed in this paper, we tested it on simulated data on four and six-taxon trees. Following the work of Schwartz and Mueller (2010), for four taxon trees we considered three sets of unrooted phylogenetic trees and fixed one inner node as the root: the set $\mathcal{T}^4_{\text{balanced}}$ corresponds to "balanced" trees with all five branches equal; the set $\mathcal{T}_{1:2}$ has the inner branch half of the length of the exterior branches; and the set $\mathcal{T}_{2:1}$ denotes a phylogenetic tree with the inner branch being the double size of the external ones (see Fig. 8.2). In $\mathcal{T}^4_{\text{balanced}}$ and $\mathcal{T}_{1:2}$ we let the length l_0 of the inner branch vary from 0.01 to 1.4, where starting from 0.05 it increases in steps of 0.05; in $\mathcal{T}_{2:1}$ we let l_0 vary in $(0, 0.7)$. For 6 taxon trees we used balanced trees $\mathcal{T}^6_{\text{balanced}}$ (see Fig. 8.2) and let the value l vary as l_0 above.

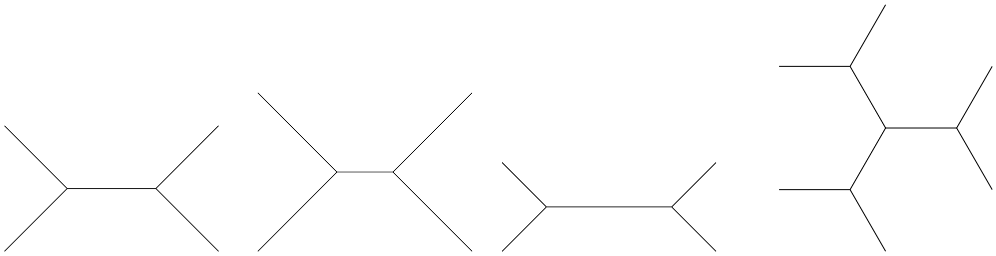


Figure 8.2: An example of $\mathcal{T}^4_{\text{balanced}}$, $\mathcal{T}_{1:2}$, $\mathcal{T}_{2:1}$ and $\mathcal{T}^6_{\text{balanced}}$ (*from left to right*).

We simulated multiple sequence alignments on trees with 4 and 6 leaves under JC69* and K80* models. We used the GenNon-h package from chapter 7 (Casanellas and Kedzierska, 2011). In brief, the program takes as an input a phylogenetic tree with given branch lengths, samples the substitution matrices corresponding to these lengths for all edges and uses them to generate the DNA multiple sequences alignment following a Markov process on the tree. The output of this software is the alignment, the substitution matrices, root distribution (whenever non-stationary) and the variances of the continuous free parameters.

Here we used alignments of length L in $\{300, 500, 1.000, 10.000\}$ for four taxa and length $1.000nt$ or $10.000nt$ for six taxa. Given an evolutionary model (JC69* or K80*), a phylogenetic tree $\mathcal{T}$ (with branch lengths), and an alignment length, we run each analysis $B = 1.000$ times. I.e., for each integer b in $1 : 1,000$, we generated substitution parameters ${}^b\xi = ({}^b\xi^e_i)_{i,e}$ for the tree $\mathcal{T}$ and a multiple sequence alignment of the corresponding length. Then we estimated the parameters using Empar.

8.3 Identifiability

An important aspect to bear in mind when conducting statistical analysis is the identifiability of the assumed model. As shown in Allman and Rhodes (2003), the GMM model, and thus all its submodels, are identifiable up to a permutation. Namely, there is a set of parameters closed under permutation of rows, which will lead to the same estimated joint probability. In practical applications this means that the matrices recovered are permuted with respect to the underlying ones. Zou et al. (2011) refer to this problem as non-indentifiability. As noted also in Zou et al. (2011), incorrect order of rows in the matrices can lead to a negative determinant of the substitution matrix in which case the branch lengths cannot be calculated.

This is in fact, the term "identifiable by rows" was coined by Allman and Rhodes (2003) and properly reflects its nature. Non-identifiability is a condition much more difficult to resolve (if at all). In the first case, we are able to recover the parameters by applying the correct permutation to the rows of the matrices, while in the latter it may not be possible. In the algorithm underlying `Empar` we focus on the biologically relevant parameters and assume the diagonal entries to be larger then the off-diagonal ones. As shown by Chang (1996), the matrices of this type form a subset of the matrices diagonalizable by rows for which the identifiability holds. However, due to the error introduced by limited data (short alignments), the labeling of the parameters may not be recovered.

We expected this problem to arise in short data sets and large branch length, as those correspond to the substitution matrices with smaller diagonal value. For all the data sets used for tests, we calculated the percentage of cases among the 1000 simulations for which the parameters estimated by the EM algorithm were permuted. This was only observed in the data sets of 300nt and 1000nt. In the first case the estimated matrices were permuted when the initial branch length was 0.55 or longer and corresponded to 0.005-0.023% of the cases. In the latter for the branches of 0.6 or longer with at most 0.001% permuted matrices. Shorter branch length and longer alignments did not suffer from the above problem and recovered the underlying order in all of the cases.

In order to ensure the reliability of the algorithm we designed a procedure that scans the tree in the search of the permutations that maximize the number of substitution matrices with larger diagonal entries. As it is not possible to maximize it for all edges, the goal is to find the permutations giving more weights to the lower parts of the tree, starting with the nodes corresponding to the outer branches. We explain this procedure in what follows. Given a tree $\mathcal{T}$ and substitution parameters A^e, for each interior node x we let $S(x)$ be the permutation of $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$ that maximizes the sum of diagonal entries on the matrices assigned to its outgoing edges after performing the corresponding permutation on their rows. Having estimated the parameters using `Empar`, we apply recursively $S(x)$ to the subtrees of $\mathcal{T}$ starting from its outer nodes towards the root.

8.4 Results and discussion

We present the results on the simulated data sets and discuss their dependence on the length of the alignments, the length of the branches and the positioning of the branches in the tree: the so-called depth of the branch (1 for external branches and 2 for internal branches in our case; Schwartz and Mueller (2010)). When there is more than one branch with the same depth, we chose one of the branches randomly (the results were the same for all branches of the same depth). We present first the results on 4-taxa as a test on the accuracy of the method and then on 6-taxa. Note that for the `JC69`* model, there is a $1-1$ correspondence between the branch length and the free parameters of the substitution matrix. However, for the `K81`* model the target distribution differs in each sample as, given branch lengths l on the edges of the tree, `GenNon-h` generates substitution matrices with the assigned branch lengths for the corresponding edges.

As a main measure of the performance of `Empar` we present the proportion of significant $p-values$ for the estimated parameters. This is based on the χ^2_d test in (8.2): for each edge we calculated the $p-value$ of the free parameters using the asymptotic results of (8.2). The $p-values$ are a measure of strength of evidence against the null hypothesis– the smaller the values, the stronger the evidence against the null hypothesis. A similar thing holds for exceptionally large $p-values$: they imply small difference between parameters and their estimates that is not to be expected by chance.Therefore, to test whether the algorithm successfully recovers the true evolutionary parameters, we presented the proportion of samples for which the $p-value$ lied in the interval $(0.05, 0.95)$. The results are shown table 8.1 for the `JC69`* model on the $\mathcal{T}_{1:2}$ tree (see remaining tables in the appendix C: Tab. C.1, and C.2 also for the `JC69`* model and Tab. C.3, C.4 and C.5 for the `K81`*). We observe that, even for short alignments of 300nt, the null hypothesis cannot be rejected in approximately 95% of the samples.

We employed a few measure that quantify the error in the estimates of the parameters and the branch lengths.

8.4.1 Estimation error

For a depth 1 and 2 branch, e, we quantified the overall divergence between the original and estimated parameters using the induced L_1 norm of the difference between the substitution matrices: $||A^e - \hat{A}^e||_1$. This norm was defined earlier in section 4.2 and used in the expression for the upper bound on the error in branch lengths estimation in the formula (4.2). Since for the `JC69`* and the `K81`* models all column are the same, the formula becomes:

$$D(\xi^e, \hat{\xi}^e) = \sum_{i=1}^{4} \mid A^e_{i,1} - \hat{A}^e_{i,1} \mid . \tag{8.4}$$

Figure 8.3 depicts the results for `JC69`* and `K80`* on the three phylogenies on four taxa for different alignment lengths. The shapes of the distribution of D for the two models in the corresponding plots are very similar. As expected, the method performs worse for large branch lengths and short alignments. There is a significant improvement with

Table 8.1: The relative frequency of the χ^2 tests based on the asymptotic normality of the maximum likelihood estimator with *p-value* $\in (0.05, 0.95)$, calculated from 1.000 simulations under the `JC69`* model. Each data set was a multiple sequence alignment of length L generated on the $\mathcal{T}_{1:2}$ tree with branch lengths set to the values indicated by the first column. *Left*: results for the depth 1 branches; *right*: results for the depth 2 branch.

	depth 1				depth 2			
l \ L	300nt	500nt	1.000nt	10.000nt	300nt	500nt	1.000nt	10.000nt
0.01	0.971	0.972	0.968	0.946	0.972	0.949	0.868	0.958
0.05	0.947	0.951	0.947	0.948	0.974	0.943	0.953	0.952
0.10	0.949	0.953	0.964	0.952	0.952	0.948	0.948	0.955
0.15	0.952	0.954	0.958	0.938	0.946	0.953	0.940	0.947
0.20	0.957	0.944	0.944	0.954	0.949	0.965	0.944	0.954
0.25	0.957	0.955	0.955	0.956	0.945	0.939	0.955	0.936
0.30	0.957	0.943	0.945	0.955	0.943	0.946	0.941	0.948
0.35	0.952	0.943	0.958	0.958	0.948	0.943	0.950	0.960
0.40	0.955	0.946	0.947	0.957	0.951	0.951	0.936	0.944
0.45	0.949	0.944	0.944	0.947	0.948	0.955	0.958	0.958
0.50	0.948	0.935	0.942	0.941	0.929	0.949	0.954	0.946
0.55	0.954	0.949	0.946	0.957	0.936	0.944	0.944	0.952
0.60	0.940	0.942	0.937	0.953	0.944	0.934	0.948	0.955
0.65	0.940	0.934	0.955	0.952	0.938	0.938	0.945	0.948
0.70	0.944	0.936	0.942	0.946	0.917	0.940	0.944	0.948
0.75	0.922	0.932	0.947	0.934	0.922	0.932	0.943	0.950
0.80	0.909	0.932	0.926	0.957	0.957	0.928	0.943	0.941
0.85	0.912	0.912	0.932	0.948	0.968	0.930	0.936	0.947
0.90	0.870	0.885	0.919	0.951	0.980	0.918	0.929	0.953
0.95	0.852	0.888	0.939	0.951	0.981	0.965	0.908	0.944
1.00	0.824	0.866	0.893	0.935	0.982	0.981	0.896	0.933
1.05	0.816	0.853	0.889	0.930	0.980	0.981	0.898	0.937
1.10	0.806	0.852	0.891	0.921	0.990	0.995	0.925	0.945
1.15	0.784	0.812	0.867	0.938	0.980	0.987	0.982	0.951
1.20	0.797	0.785	0.823	0.923	0.986	0.986	0.984	0.942
1.25	0.786	0.803	0.824	0.938	0.983	0.981	0.984	0.941
1.30	0.789	0.793	0.800	0.894	0.981	0.976	0.992	0.925
1.35	0.755	0.787	0.786	0.893	0.973	0.991	0.989	0.912
1.40	0.761	0.789	0.785	0.864	0.970	0.974	0.994	0.879

the increase in the alignment length. For 10.000nt the estimates of the parameters were very accurate. The performance under the `JC69`* model (Fig. 8.3(a)) is better than that of `K80`* (Fig. 8.3(b)) for shorter branch lengths.

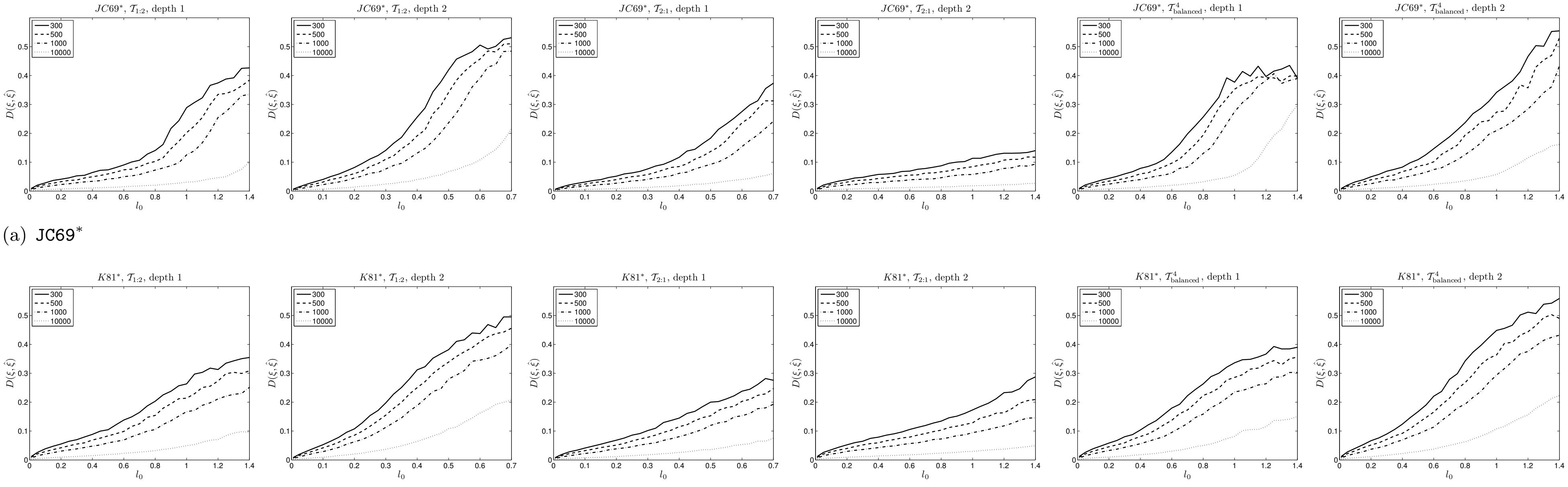
8.4.2 Parameter dispersion

Figure 8.4 shows the variances of the estimated parameters for depth 1 and 2 branches on the $\mathcal{T}_{\text{balanced}}$, $\mathcal{T}_{1:2}$, $\mathcal{T}_{2:1}$ trees under the `JC69`* model. The variances show an exponential increase, most significant for the $\mathcal{T}^4_{\text{balanced}}$ tree, both depths of the branches and the depth 2 branch of $\mathcal{T}_{1:2}$. The results for the depth 1 branch in $\mathcal{T}_{\text{balanced}}$ and $\mathcal{T}_{1:2}$ are very similar. The smallest variance was observed for the depth 2 of $\mathcal{T}_{2:1}$. We observe that for alignments of length $10.000nt$ we can say that the method is quite accurate. The reader can find these results in the appendix C (see Tab. C.6, C.7 and C.8).

For the `K81`* model we summarized the results on variances for each edge as the mean of combined variances of all samples (see Fig. 8.5). The results are analogous to those of the `JC69`* model. As expected, the variances are less dispersed and lower for shorter branches and longer alignments (see Tab. C.9, C.10, C.11 in the appendix C).

8.4.3 Error in the branch lengths

Using the formula (1.2) we calculated the actual difference between the branch length l_0 computed from the original parameters ξ and the branch length $\hat{l}$ computed using their MLEs $\hat{\xi^e}$. Negative values of this score imply overestimation of the branch length, while positive values indicate underestimation. The results are shown in Figures 8.6 and 8.7.



(a) `JC69`*

(b) `K81`*

Figure 8.3: Divergence $D(\xi, \hat{\xi})$ between the parameters, ξ, and their estimates, $\hat{\xi}$, calculated by `Empar`. Horizontal axis: original length of the inner branch.

In the case of JC69* we observe that the method presented here does not tend to underestimate or overestimate the lengths for the depth 1 branches in all the 4-taxon trees ($l_0 - \hat{l}$ is centered at 0, see Fig. 8.6). The depth 2 branches have a tendency towards overestimation of the length for branches longer than ≈ 0.45 for $\mathcal{T}_{1:2}$, ≈ 0.9 for $\mathcal{T}_{2:1}$ and ≈ 0.8 for the $\mathcal{T}^4_{\text{balanced}}$ trees. In the latter case, lengths longer then 1.2 for alignments up to 1.000nt show opposite trend of underestimating the true lengths. The values were accurate when the alignment lengths were increased in the case of $\mathcal{T}_{1:2}$ and $\mathcal{T}_{2:1}$. On the other hand, for $\mathcal{T}^4_{\text{balanced}}$ the alignments of 10.000nt resulted in overestimation.

In the K81* models the results are significantly more accurate (see Fig. 8.7). There is a trend of underestimation for branches longer than ≈ 0.9 for shorter alignments. That is especially noticable for $\mathcal{T}^4_{\text{balanced}}$ and depth 1 branches of $\mathcal{T}_{1:2}$. This trend diminishes with an increase in the alignment length.

Overall, in the case of both models, the variance of the estimate is small for shorter lengths and both depth 1 and 2 branches of the $\mathcal{T}_{2:1}$ tree.

We also calculated the tree length (i.e. the sum of its branch lengths) from the estimated parameters and compared it to the theoretical result on the original branch length l_0: $4.5l_0$ for $\mathcal{T}_{1:2}$ (l_0 depth 1 branch), $3l_0$ for $\mathcal{T}_{2:1}$ (l_0 for depth 2 branch) and $5l_0$ for $\mathcal{T}^4_{\text{balanced}}$. The rightmost columns of Figures 8.6 and 8.7 show the results for 4-taxon trees for the JC69* and K81* models. The length of the tree is estimated accurately for all trees, the estimates being best for $\mathcal{T}_{2:1}$. The variance is small and decreasing with an increase in the data length. As the sequences get longer, the distribution is centered around the true value. This is especially visible for the K81* model (see Fig. 8.7).

8.4.4 Results for larger trees

We increased the number of species and run the analysis on the 6-taxon balanced tree, $\mathcal{T}^6_{\text{balanced}}$, under the K81* model, $\mathbf{l} \in \{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.4\}$ and the alignment lengths of 1.000nt and 10.000nt. The $p - values$ of the corresponding tests confirm that the performance of the algorithm is very satisfactory (see Tab. C.12 in the appendix C). We have seen in the 4-taxon study that the tree with equal branch lengths gave the worst results than the unbalanced trees. Thus, we expect the results of the depth 2 branch to be similarly challenged.

Figure 8.8 depicts the estimated tree lengths. It can be seen that the estimates are accurate and the results improved for the alignments of 10.000nt. As expected, the variance of the estimates increases with the increase in the length of the branch. By formula 1.2, long branches correspond to small values of the determinant of the transition matrix. Thus, statistical fluctuations in the parameter estimates have greater impact on the resulting length of the tree.

Next, we calculated the difference between the oryginal and estimated branch lengths. In Figure 8.9(a), we see that the depth 1 branches show some degree of underestimation of the length for lengths $1.1 - 1.4$ and alignments of 1.000nt. In the case of 10.000nt, the results improve and can be expected to show little bias for even longer data sets. Branches of depth 2 show higher degree of underestimation with improve-

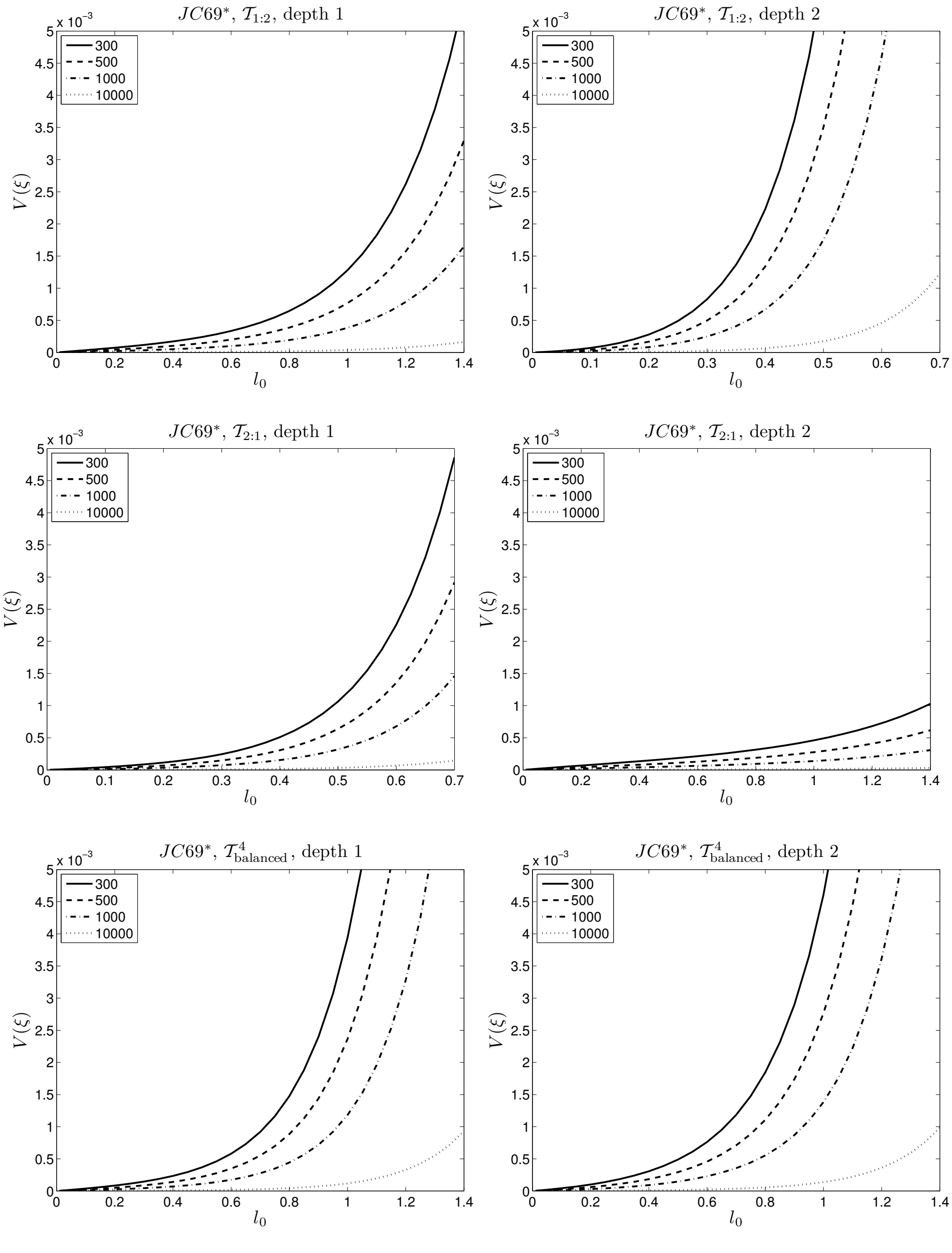


Figure 8.4: Distribution of variances of the estimated parameters for different alignment lengths and different lengths of the depth 1 (*left*) and depth 2 (*right*) branches under the `JC69`* model: $\mathcal{T}_{1:2}$ (*top*), $\mathcal{T}_{2:1}$ (*middle*), $\mathcal{T}^4_{\text{balanced}}$ (*bottom*).

ment for longer data set. The estimation error of the parameters given in the formula 8.4 is shown in Figure 8.9(b). For branches of depth 1 and the data of length 10.000 the distance is ≈ 0.2. In the case of branches of depth 2, it is almost doubled for both alignment lengths. In both cases, branch lengths up to 0.5 give satisfactory results. The error of the estimates for longer branches seems to be approaching a plateau.

Combined variance of the estimated parameteres is much decreased for the 10.000nt data sets in comparison with the 1.000nt, and is smaller for the depth 1 branch (see Fig.8.9(c)). Again, the exponential shape of the plot can be attributed to the logarithm appearing in the formula 1.2

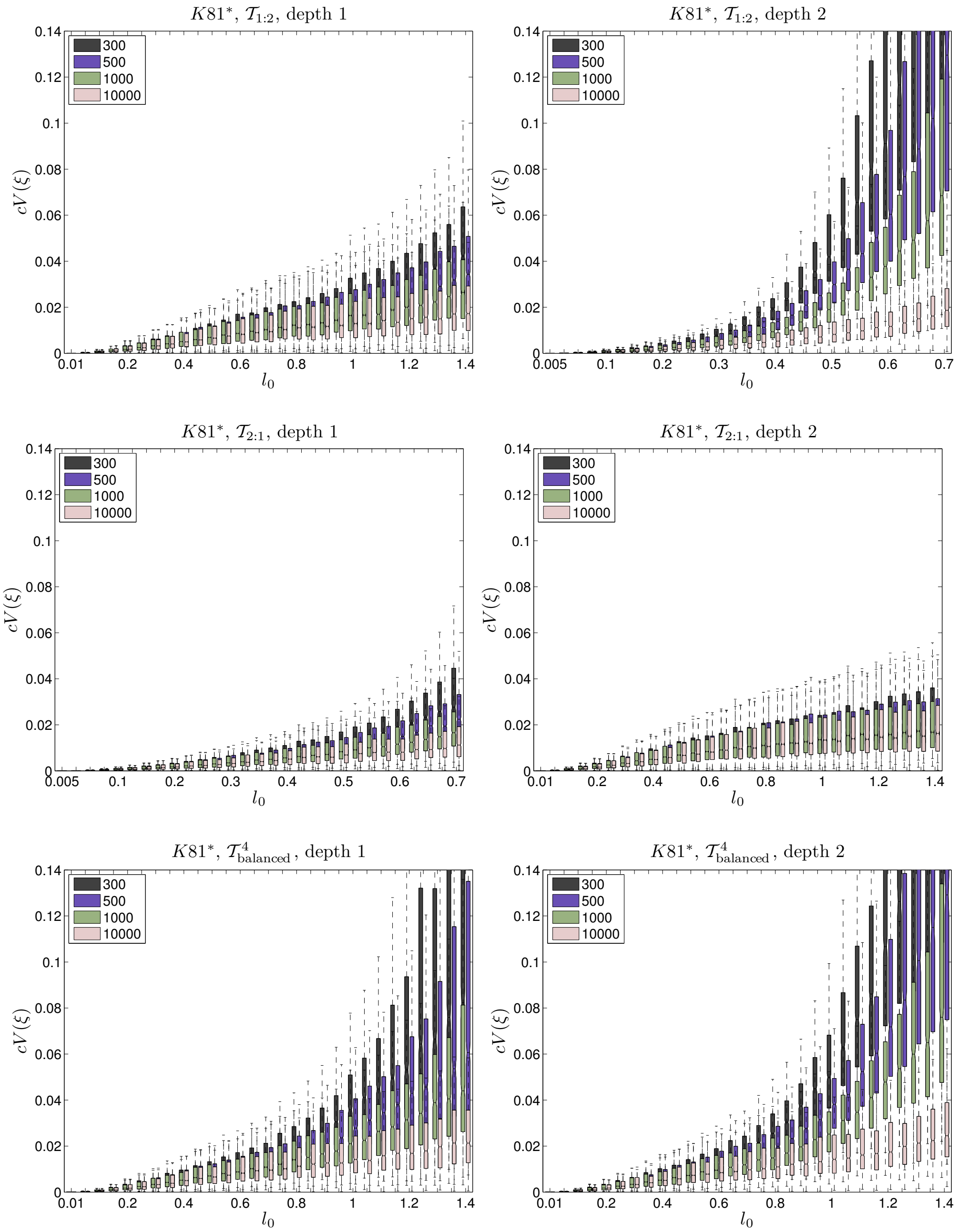


Figure 8.5: Distribution of combined variances of the estimated parameters for different alignment lengths and different lengths of the depth 1 (*left*) and depth 2 (*right*) branches under the `K81`* model: $\mathcal{T}_{1:2}$ (*top*), $\mathcal{T}_{2:1}$ (*middle*), $\mathcal{T}^4_{\text{balanced}}$ (*bottom*).

To summarize, we evaluated the performance of the EM algorithm for phylogenetic parameter estimation under various circumstances on simulated data sets. As expected, `Empar` performs best for long alignments and short branch lengths. Also, the results are better for less complex models due to the smaller number of parameters to be estimated.

It is worth noticing that even for short alignments of 300nt or 500nt on 4 taxa, the null hypothesis "estimated parameters are equal to the original parameters" couldn't be rejected in approximately 95% of the cases. Moreover, the estimation of branch lengths is very accurate even for such short alignments.

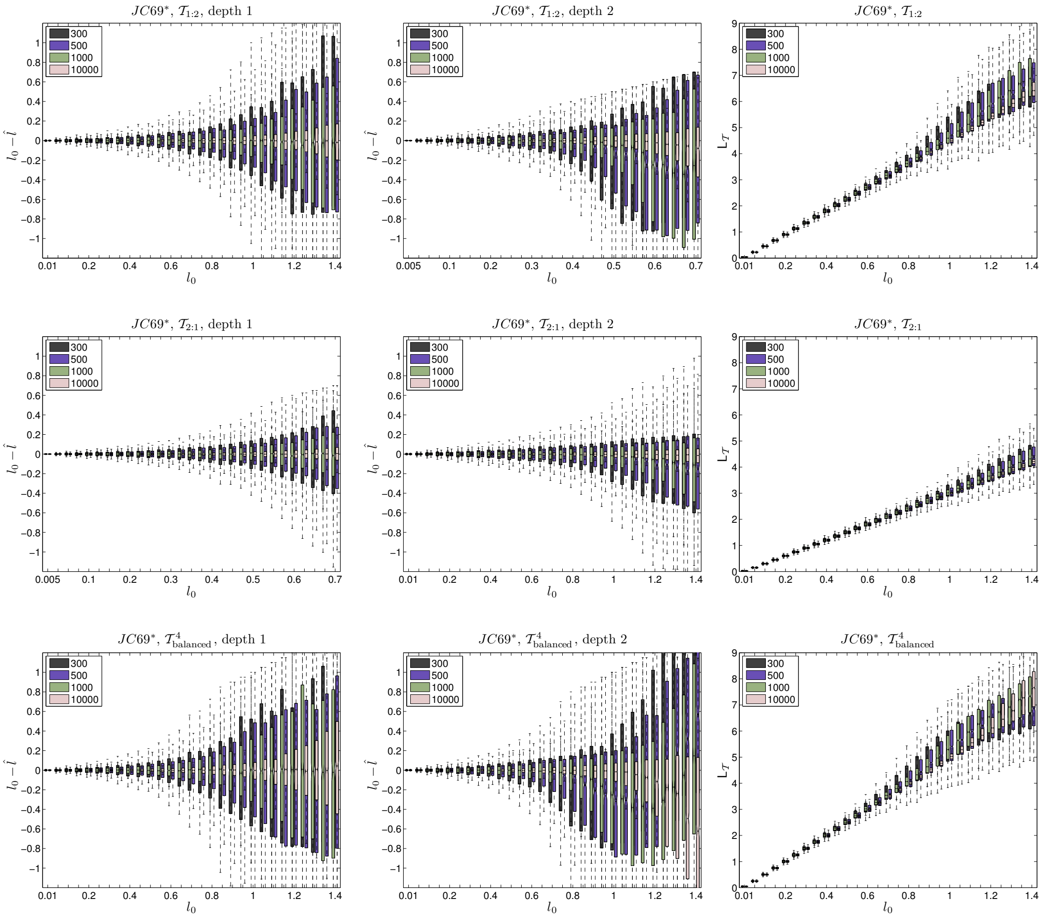


Figure 8.6: Error in the branch length estimation measured as the difference between the initial and the estimated branch lengths, $l_0 - \hat{l}$, in the 1.000 simulated data sets along the $\mathcal{T}^4_{\text{balanced}}, \mathcal{T}_{1:2}, \mathcal{T}_{2:1}$ trees under the JC69* model (*left and middle columns*). *Rightmost column* displays the distribution of the estimated length of the tree, where l_0 labelling the horizontal axis corresponds to the length of the internal branch in $\mathcal{T}$.

Table 8.2: Time it took for Empar to estimate parameters of alignments of 1 and 10 thousands of nucleotides, generated on star trees with varying number of nodes, n, and equal branch length of 0.5.

length (nt)\ n	3	4	5	6	7	8
1.000	0.004	0.02	0.033	0.222	1.049	7.14
10.000	0	0.011	0.043	0.171	1.044	6.95

There are two drawbacks to the method. Firstly, there is an exponential increase in the computational time with the increase in the number of taxa. This is computational limitation due to the fact that the algorithm computes large matrices of dimension exponential in the total number of nodes of a tree. Running time of Empar on star trees with 3-8 nodes and equal branches of 0.5 on Ubuntu 11.10, Intel Core i7 920 at 2.67 GHz with 6 Gb is given in Table 8.2. Secondly, the memory usage of Empar is approx. $8 * 4^{|\text{nodes}|}$, which corresponds to the memory footprint of the matrix in the EM algorithm. For exaple, for this matrix to fit in the memory of a 6Gb machine, we get the bound on the number of nodes: $|\text{nodes}| \leqslant 14$.

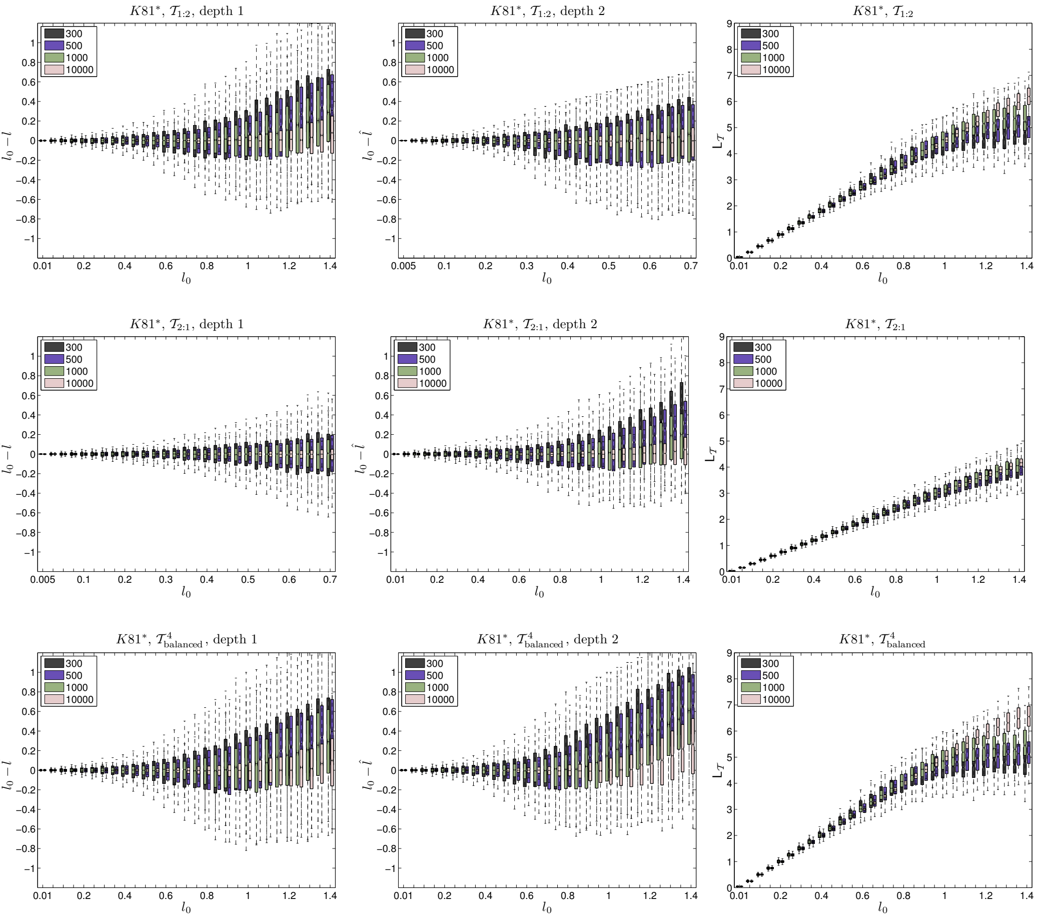


Figure 8.7: Error in the branch length estimation under the K81* model (see Fig. 8.6 for details).

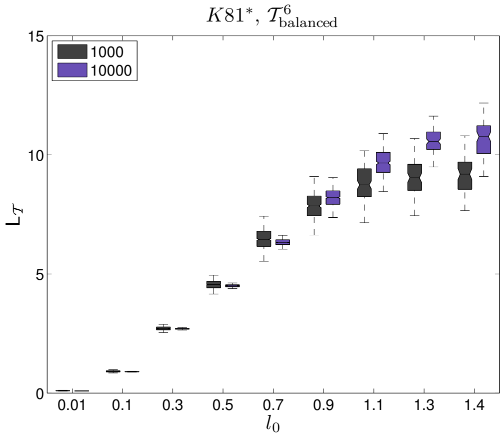
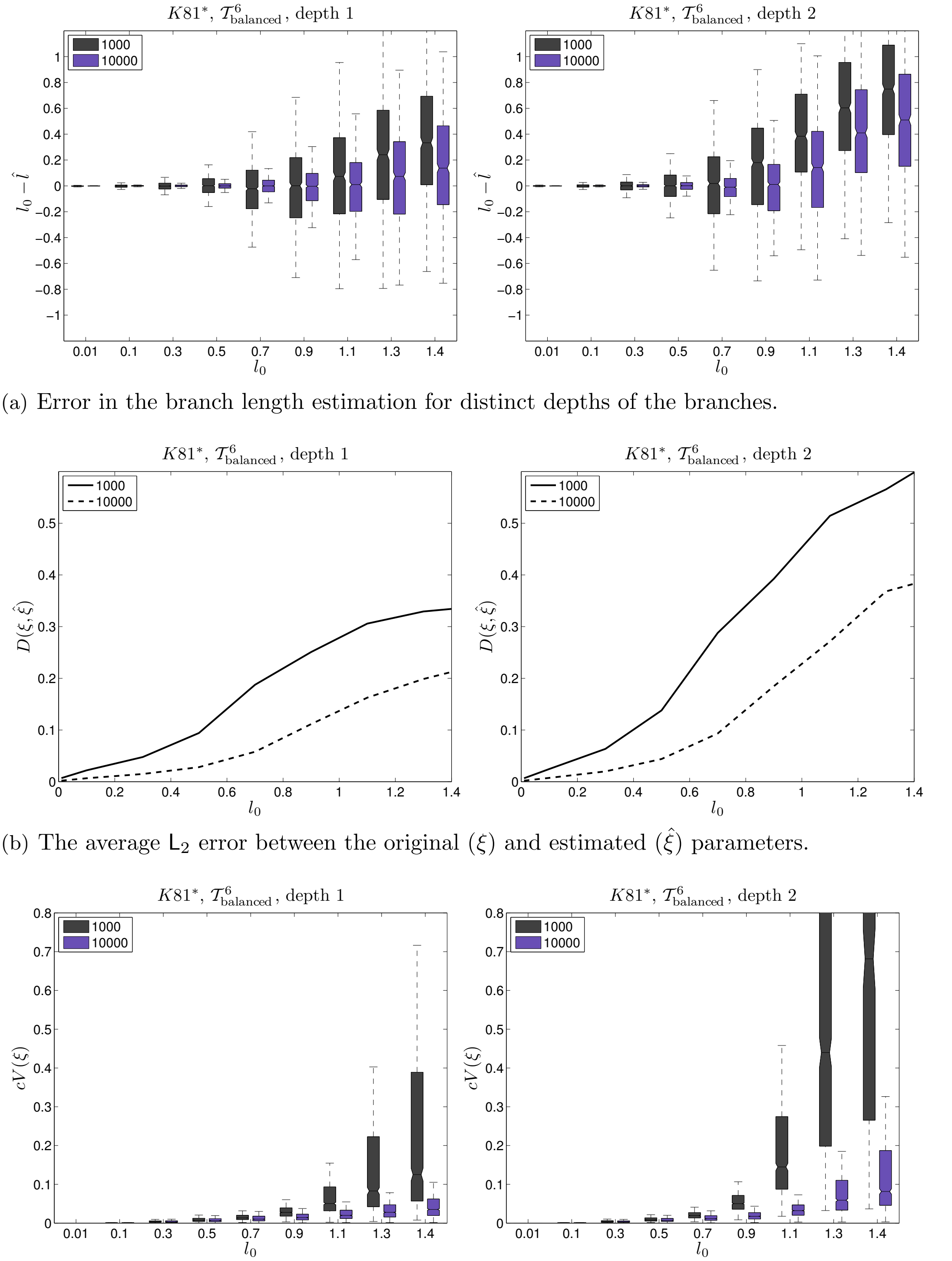


Figure 8.8: Estimated tree length as a function of the initial length of a branch of $\mathcal{T}^6_{\text{balanced}}$ ($\mathsf{L}_{\mathcal{T}} = 9l_0$) in 1.000 data sets generated under the K81* model.

To sum up, we suggest Empar as a highly reliable method for estimating branch lengths for a small number of taxa on trees of short branch lengths, even for short alignment.

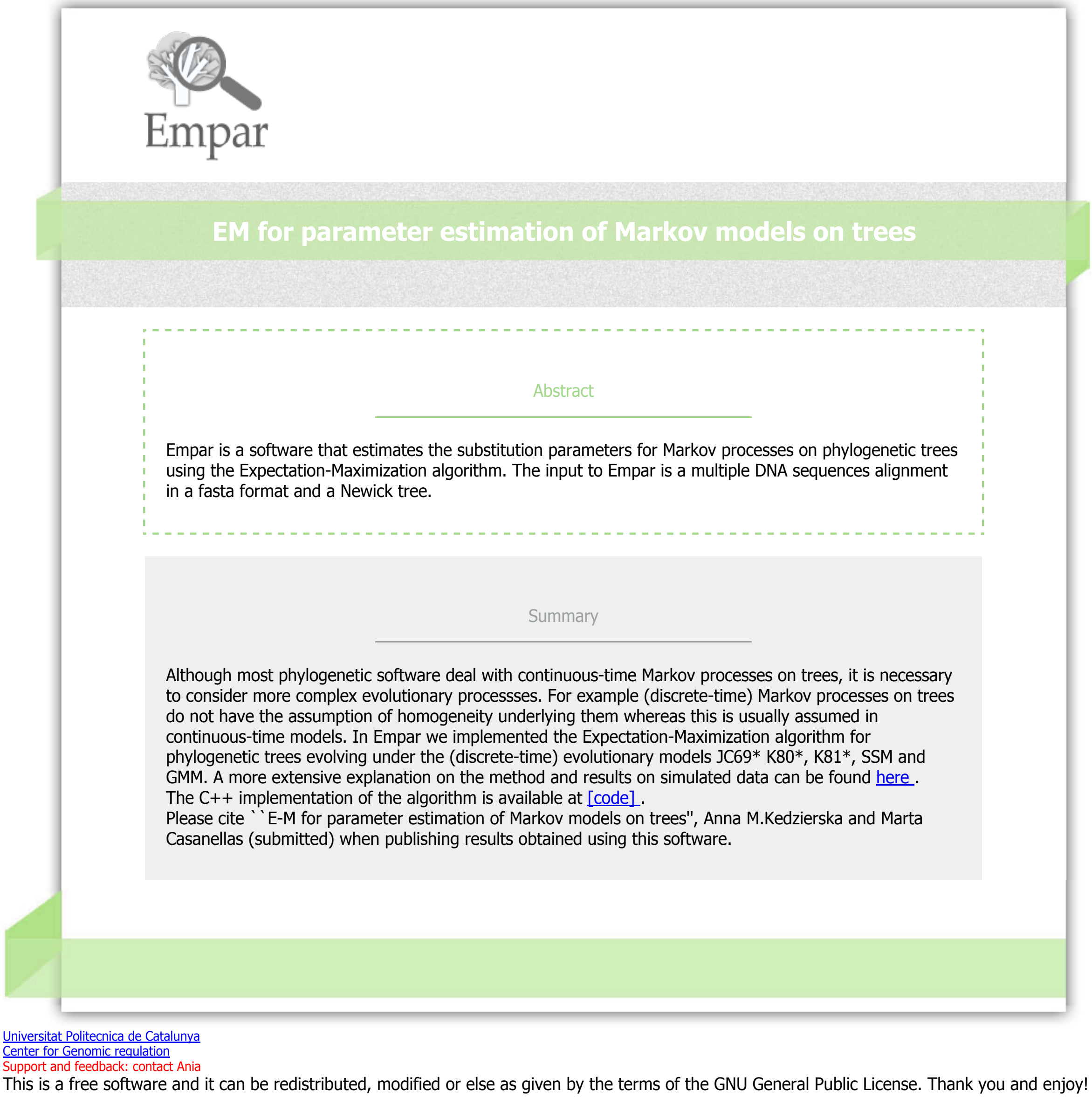


(a) Error in the branch length estimation for distinct depths of the branches.

(b) The average L_2 error between the original (ξ) and estimated ($\hat{\xi}$) parameters.

(c) Distribution of the combined variance for distinct depths of the branches.

Figure 8.9: Results for the 1.000 data sets generated on the $\mathcal{T}^6_{\text{balanced}}$ tree for the $\texttt{K81}^*$ model.



EM for parameter estimation of Markov models on trees

Abstract

Empar is a software that estimates the substitution parameters for Markov processes on phylogenetic trees using the Expectation-Maximization algorithm. The input to Empar is a multiple DNA sequences alignment in a fasta format and a Newick tree.

Summary

Although most phylogenetic software deal with continuous-time Markov processes on trees, it is necessary to consider more complex evolutionary processses. For example (discrete-time) Markov processes on trees do not have the assumption of homogeneity underlying them whereas this is usually assumed in continuous-time models. In Empar we implemented the Expectation-Maximization algorithm for phylogenetic trees evolving under the (discrete-time) evolutionary models JC69* K80*, K81*, SSM and GMM. A more extensive explanation on the method and results on simulated data can be found here .
The C++ implementation of the algorithm is available at [code] .
Please cite ``E-M for parameter estimation of Markov models on trees'', Anna M.Kedzierska and Marta Casanellas (submitted) when publishing results obtained using this software.

Universitat Politecnica de Catalunya
Center for Genomic regulation
Support and feedback: contact Ania
This is a free software and it can be redistributed, modified or else as given by the terms of the GNU General Public License. Thank you and enjoy!

Chapter 9

SPIn: Model Selection in Phylogenetics based on linear INvariants

This chapter is a collaboration with Marta Casanellas, Mathias Drton and Roderic Guigó.

Model selection in phylogenetics is a challenging problem. Even more so, if one considers phylogenetic mixtures.

Specification of an evolutionary model of suitable complexity for the nucleotide substitution process at hand is often viewed as a 'pre-inference' step in phylogenetic analysis. However, as has been emphasized in the literature (Posada and Crandall, 2001; Ripplinger and Sullivan, 2008), this step should be addressed with care as it can strongly impact the accuracy of the reconstructed topology and the estimates of the branch length. Inference of an appropriate evolutionary model is further challenged when the data evolved under a nonhomogeneous model (rate matrices vary across the edges) or along multiple trees (phylogenetic mixture) (Hillis et al., 1994; Ho and Jermiin, 2004; Lockhart et al., 1996; Sullivan and Swofford, 2001; Swofford et al., 2001; Bruno and Halpern, 1999; Kolaczkowski and Thornton, 2004).

Ripplinger and Sullivan (2010) show that the performance of established model selection methods depends highly on the underlying tree topology. A common practice, however, adopts a circular argument: the tree and the parameters of interests are estimated by choosing a model supported by a pre-computed tree (e.g., the neighbor-joining tree based on Jukes-Cantor distances). Moreover, as outlined above, available methods for selecting a model of evolution typically assume constant rate parameters at each point in time as well as a single tree topology underlying the data-generating process (e.g. Foster, 2004; Huelsenbeck et al., 2004; Posada, 2008). Mossel and Vigoda (2005) and Ronquist et al. (2006) discuss poor mixing of the phylogenetic Markov chain Monte Carlo (MCMC) in the presence of mixed phylogenetic signals. We propose an approach designed to deal with both nonhomogeneous and mixed data with no a priori requirement of a tree topology.

As pointed out by Fu and Li (1992b), Steel et al. (1992) and Felsenstein (2003), *model invariants* could potentially be used to discriminate between different models of base change. Following on the results introduced in the prior sections, we introduce a method for model Selection in Phylogenetics based on linear INvariants (SPIn), which uses recent insights on linear invariants to characterize a model of nucleotide evolution for phylogenetic mixtures on any number of components.

In addition, for a given model and a number of sequences, `SPIn` calculates the maximum number of trees to be considered in a mixture. As proved in Section 6.4, mixture models with more components than a particular bound cease to be identifiable. The outcomes of presented in this Section were published in Kedzierska et al. (2012).

Remark 9.1. The sets of equations provided in Theorem 6.11 describe the linear spaces of dimensions that are exponential in n. However, for its biological application one does not need to consider all the equations but only those containing the patterns observed in the data (in real applications the number of different columns in an alignment is really small compared to the dimension of these spaces).

Selecting a model based on biological data requires a statistical assessment of the vanishing of the linear invariants for each model. Let $\mathcal{L}^{\mathcal{M}}$ be the linear space formed by all distributions satisfying the linear invariants for the model $\mathcal{M}$ (see Def. 6.6). For the models considered here, $\mathcal{L}^{\mathcal{M}}$ is defined by equalities among pairs of entries of $\mathbf{p}_T^{\mathcal{M}}(\theta)$ (see Thm. 6.7), where θ denotes the set of model parameters. Hence, the maximum likelihood estimate is unique, that is, given data D there exists a unique point $\hat{\theta} \in \mathcal{L}^{\mathcal{M}}$ for which the likelihood function $L(\theta, \mathcal{M}) = Prob(D \mid \theta, \mathcal{M})$ attains its maximum for $\theta \in \mathcal{L}^{\mathcal{M}}$. To score the models, we use a variant of the AIC which includes a small sample correction along with the penalty for model complexity:

$$AIC_c = -2\log(L(\hat{\theta}, \mathcal{M})) + 2d + \frac{2d(d+1)}{L-d-1},$$

where L is the sample size (alignment length) and d is the dimension of the linear space $\mathcal{L}^{\mathcal{M}}$. In Proposition 6.9 we explicitly the dimension of the $\mathcal{L}^{\mathcal{M}}$ for the equivariant models. The number of invariants for each model is 4^n minus its dimension.

The model selected by `SPIn` is the one that minimizes AIC_c. For ranking purposes, the output of the algorithm includes the ratios of normalized Akaike weights

$$w_i = \frac{e^{-\frac{1}{2}\Delta_i}}{\sum_i e^{-\frac{1}{2}\Delta_i}}, \quad \Delta_i = AIC_{c,i} - \min_j(AIC_{c,j})$$

and $AIC_{c,i}$ is the AIC_c score of a model $\mathcal{M}_i$.

`SPIn` is a C++ package available at

`http://genome.crg.es/cgi-bin/phylo_mod_sel/AlgModelSelection.pl.`

We tested `SPIn` on synthetic data on trees of 4 OTUs following the guidelines of Posada and Crandall (2001). The simulations were done for a wide range of parameters in the continuous-time homogeneous and discrete-time nonhomogeneous settings, for a single tree topology and along a mixture of two distributions both on the same and different tree topologies. Though at this point the existing software packages such as *jModelTest* (Posada, 2008), *PAML* (Yang, 2007), *Phylip* (Felsenstein, 1993) or *PhyML* (Guindon and Gascuel, 2003) offer a larger selection of models than those included in

SPIn, these methods are not consistent for phylogenetic tree mixtures. For instance, the models considered by these methods do not allow mixtures of distinct tree topologies. We demonstrate this in the Results section, where we evaluate the performance of *jModelTest*. Recently, Nguyen et al. (2011) used the joint patterns at the leaves to assess the fit of an inferred model and a tree to the data. In order to show that SPIn is not biased towards over-complex models, we have analyzed one of the data sets used in Nguyen et al. (2011) (see Discussion section below).

9.1 Data

In order to assess the performance of SPIn in recovering the underlying model from $\{$JC69*, K80*,K81*, SSM$^*\}$, we simulated multiple sequence alignments on an unrooted quartet tree following the design of Posada and Crandall (2001). Specifically, we used the quartet tree space proposed by Huelsenbeck (1995), which is defined by a pair of branch-length parameters (a, b), where a determines the length of the internal branch and two peripheral branches taken from different clades, and b gives the length of the two remaining branches. Parameters a and b, representing the expected number of substitutions per site, were varied from 0.01 to 0.75 in increments of 0.02 (see Fig. 9.1).

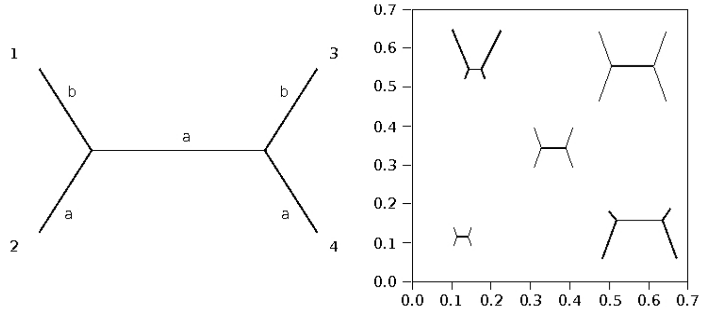


Figure 9.1: Quartet tree parameter space used for simulations (see Huelsenbeck (1995)).

We simulated 100 gap-free multiple sequence alignments of 300, 1,000 and 3,000 sites for every point (a, b) on the grid. The alignments were generated either under a single tree topology or mixtures of two trees (see below). We then computed the fraction of alignments for which the true model with a minimal sets of parameters was selected from the pool of candidate models. In graphical displays a point (a, b) is colored black if there was a 100% successful recovery. White points on the grid correspond to a 0% recovery and the values in between the two extrema are represented in a grey scale.

We used the *evolver* program from the package *PAML* (Yang, 2007) to generate the data under the continuous-time homogeneous JC69 and K80 models. We assumed a transition transversion ratio of 2 for K80 ($\kappa = 4$). In order to generate the data under the discrete hidden Markov process, we used the package introduced in Section4. As defined in (1.2), the length of a given edge is directly related to its assigned substitution matrix and is given by $\mathbf{l} = -\frac{1}{4}\log\det(A^e)$. We simulated data using an earlier Matlab version of the GenNon-h package created for this purpose as introduced in Section **??**.

We performed a number of tests on the data simulated under different parameter and model choices. Here we present the outcome of the tests and comparison of the performance of SPIn to that of the *jModelTest*.

9.2 Results

9.2.1 Single tree

We generated data on a single 4-taxon tree topology and the tree space as defined above. The resulting set of data-generating distributions is denoted by ST. The results of running SPIn under the JC69 and K80 models are shown in Figure 9.2(a). It can be seen that already for alignments as short as $300nt$, the recovery is close to perfect across the entire the tree space.

The average recovery for $300nt$ alignments was 99.9% and 97.7% and improved to 99.7% and 99.8% for length $1,000$; see Table 9.1(a). Figure 9.2(c) shows the recovery of the discrete-time JC69*, K80* and K81* models also to be high even for short alignments. The average recovery taken over the tree space and alignments of length $1,000$ was 99.7%, 96.5% and 96.8% for JC69*, K80* and K81*, respectively (see Table 9.1(b)).

9.2.2 Two tree mixtures

For the purpose of testing model recovery using SPIn on phylogenetic mixtures, we considered 2-tree mixtures on both the same and different quartet tree topologies.

First, we generated continuous-time mixture data on the same tree topology by allowing 2 gamma classes in the *evolver* package from *PAML*. The pattern of model recovery under the JC69 and K80 along these 2-tree mixtures is almost identical to that for a single tree; see Table 9.1(a).

Next, we tested the performance on 2-tree mixture data under the discrete-time hidden Markov models JC69*, K80* and K81*. Multiple sequence alignments were simulated by choosing a pair of tree topologies on 4 sequences, τ_1 and τ_2, with branch lengths fixed for τ_1 and the branch lengths of τ_2 varying over the tree space described above. We denote by MST (*mixture on the same topology*) the data-generating distributions obtained by assuming the same tree topology $\tau_1 = \tau_2$ and by MDT (*mixture on distinct topologies*) the distributions given by two different topologies $\tau_1 \neq \tau_2$. We considered two sets of branch lengths for τ_1 in the MST and MDT data sets:

(1) 0.11 for the inner branch length and two opposite peripheral branches, 0.61 for the remaining branches with a fraction of $\lambda = 0.3$ sites evolving on τ_1 (0.7 evolved on τ_2). This selection comprises the MST_1 and MDT_1 data sets.

(2) 0.31 for the inner branch length and two opposite peripheral branches, 0.41 for the remaining branches with a fraction of $\lambda = 0.5$ randomly selected sites coming from the alignment evolved on τ_1. The corresponding data sets are denoted by MST_2 and MDT_2.

In concordance with the single tree case, the recovery of the JC69* model for the MST data exceeds 99% for alignments as short as $300nt$, irrespective of the choice of

Table 9.1: Average recovery rate of the continuous-time (`JC69`, `K80`) and discrete-time models (`JC69`*, `K80`* and `K81`*) across the quartet tree space (a, b). ST: single tree under continuous and discrete-time models; Γ: single tree under continuous-time model with 2 gamma rates; MST_1, MST_2: 2-tree mixture on the same topology under discrete-time models; MDT_1, MDT_2: 2-tree mixture on different topologies under discrete-time models (see Results).

(a) `SPIn`

Model	`JC69`	`JC69`	`K80`	`K80`
Length	300	1,000	300	1,000
ST	0.999	0.997	0.977	0.998
Γ	0.999	0.998	0.940	0.998

(b) `SPIn`

Model	`JC69`*	`JC69`*	`K80`*	`K80`*	`K81`*	`K81`*
Length	300	1,000	300	1,000	300	1,000
ST	0.999	0.997	0.684	0.965	0.561	0.968

(c) `SPIn`

	Length	`JC69`*	`K80`*	`K81`*
MST_1	300	0.999	0.538	0.470
MST_2	300	0.998	0.478	0.370
MST_1	1000	0.997	0.935	0.590
MST_2	1000	0.997	0.929	0.965
MST_1	3000	0.997	0.994	0.999
MST_2	3000	0.994	0.993	0.998
MDT_1	300	0.999	0.575	0.492
MDT_2	300	0.999	0.502	0.379
MDT_1	1000	0.997	0.957	0.984
MDT_2	1000	0.998	0.952	0.977
MDT_1	3000	0.996	0.997	0.999
MDT_2	3000	0.998	0.997	0.999

(d) *jModelTest*

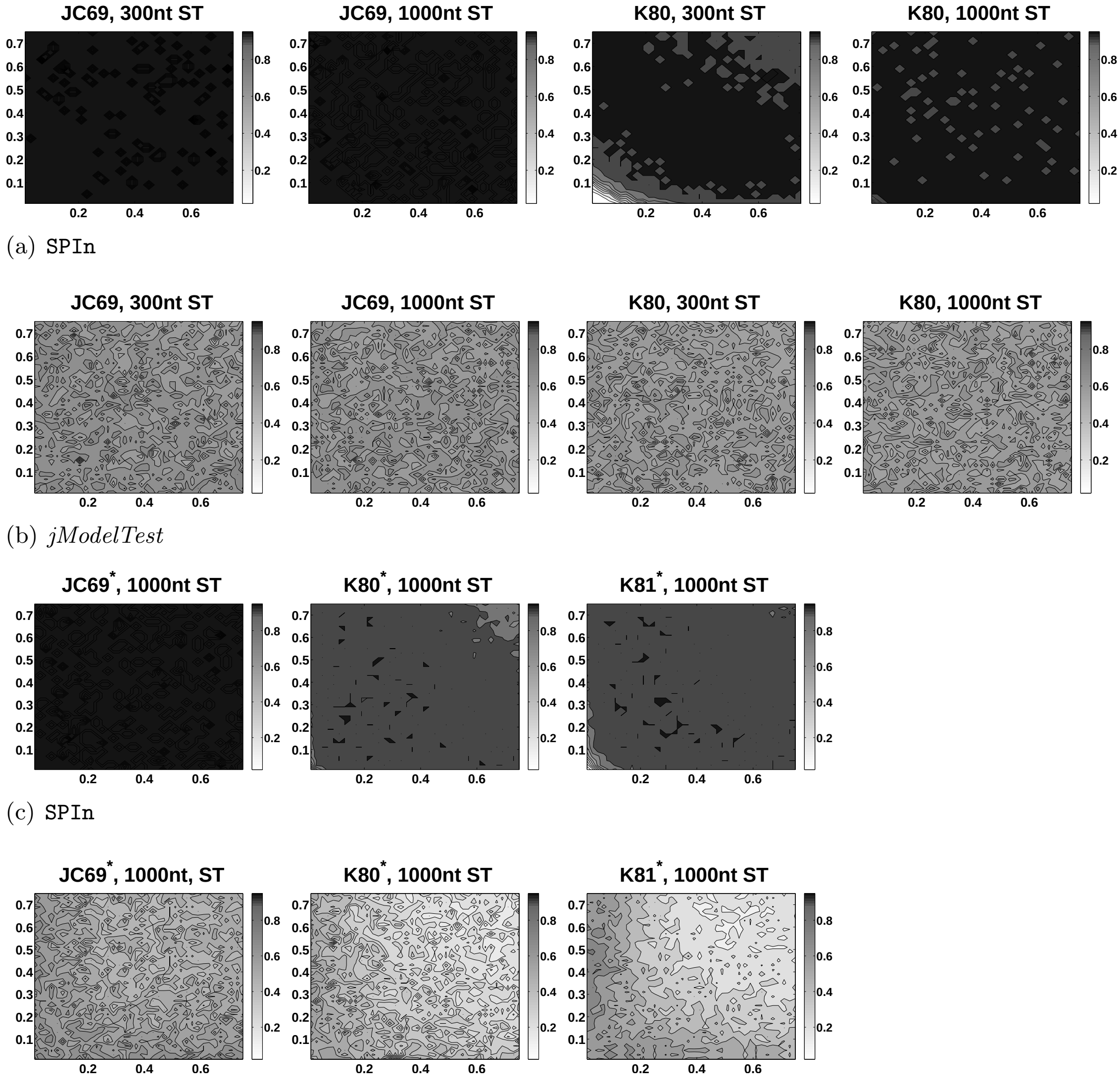
Model	`JC69`*		`K80`		`JC69`*	`K80`*	`K81`*
Length	300	1000	300	1,000	1,000	1,000	1,000
ST	0.666	0.653	0.629	0.624	0.564	0.375	0.493

Model	`JC69`* MST_2		`K81`* MST_2		`JC69`* MDT_1	`JC69`* MDT_2	
Length	300	1,000	300	1,000	3,000	300	3,000
	0.672	0.556	0.411	0.386	0.448	0.649	0.451

the parameters. See Figure 9.5(a) and Table 9.1(c) for the results on $300nt$ and $1,000nt$. As expected, it remained true for the MDT data (Figure 9.3(c), Table 9.1(c)), where the model was correctly identified at the 99% level in all data sets: $300nt$ simulated for MDT_2 and $3,000nt$ for both MDT_1 and MDT_2. At length $300nt$ the `K80`* model was recovered on average in 54% of the cases for the MST_1 (see Fig. 9.6(c)) and 48% of the cases for the MST_2 data set. Similarly lowered is the performance for the `K81`* at the alignment length of 300: 47% for the MST_1 and 37% for the MST_2 (see Fig. 9.6(c)).

The reason for this relatively low performance is the high number of parameters allowed in the (*) models due to the non-homogeneity assumption. Thus longer sequence alignments are required when using the AIC_c criterion.

For all models and their parameter choices, the recovery exceeded 99% when the alignment length was $3,000nt$ (see Fig. 9.5 and 9.6).



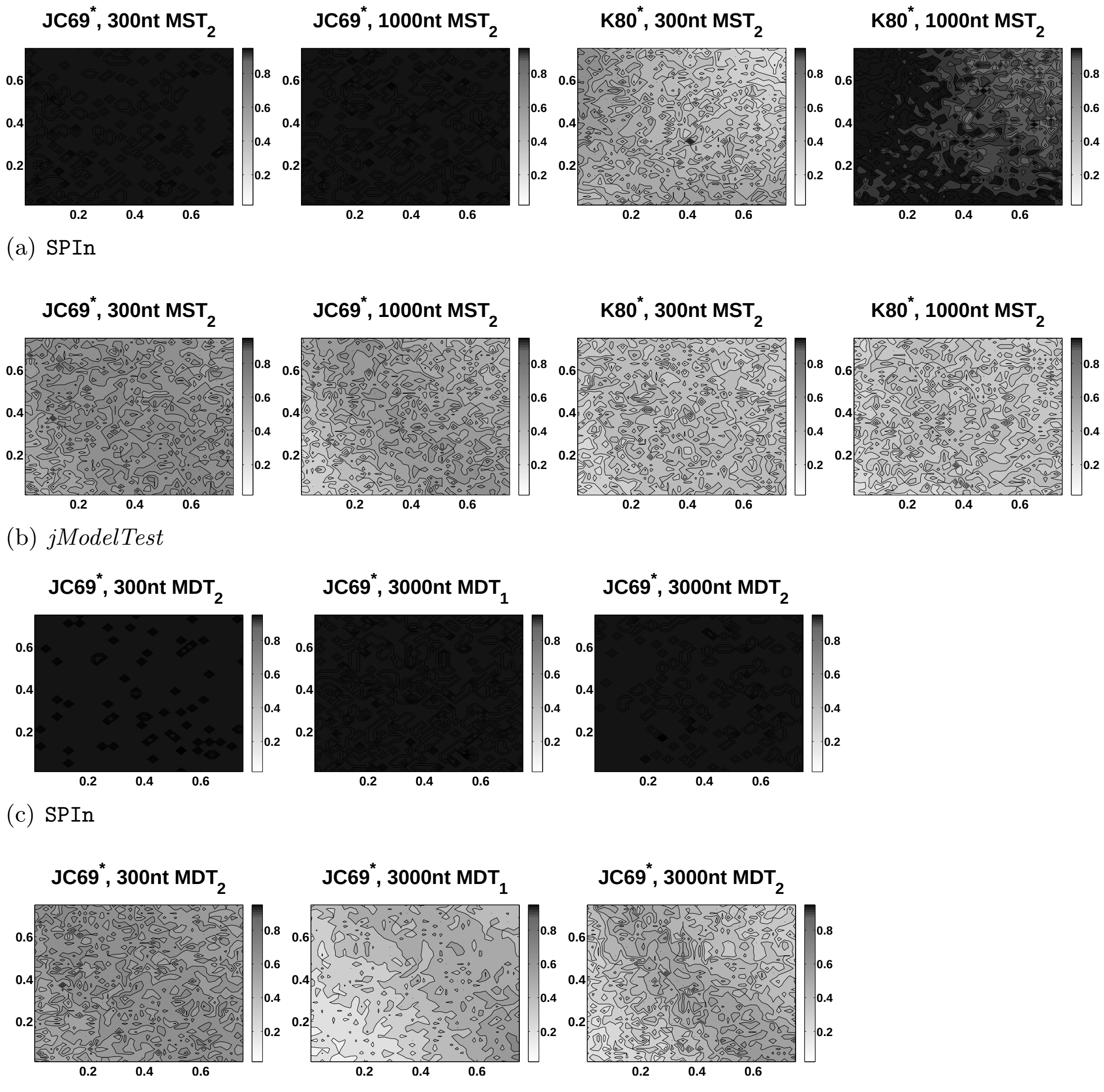
(a) `SPIn`

(b) *jModelTest*

(c) `SPIn`

(d) *jModelTest*

Figure 9.2: Plots of the fraction of correctly identified models for multiple sequence alignments of length 300 or $1,000$ generated on a single quartet tree (ST) under `JC69`, `K80`, `K81`, `JC69`*, `K80`* and `K81`*; `SPIn`: (a), (c); *jModelTest*: (b), (d). The parameters vary in the quartet tree space: (a, b) of Huelsenbeck (1995). Fractions are displayed in grey-scale ranging from 0% in white to 100% in black. Corresponding average recovery rates are given in Table 9.1(a) and (b).



(a) SPIn

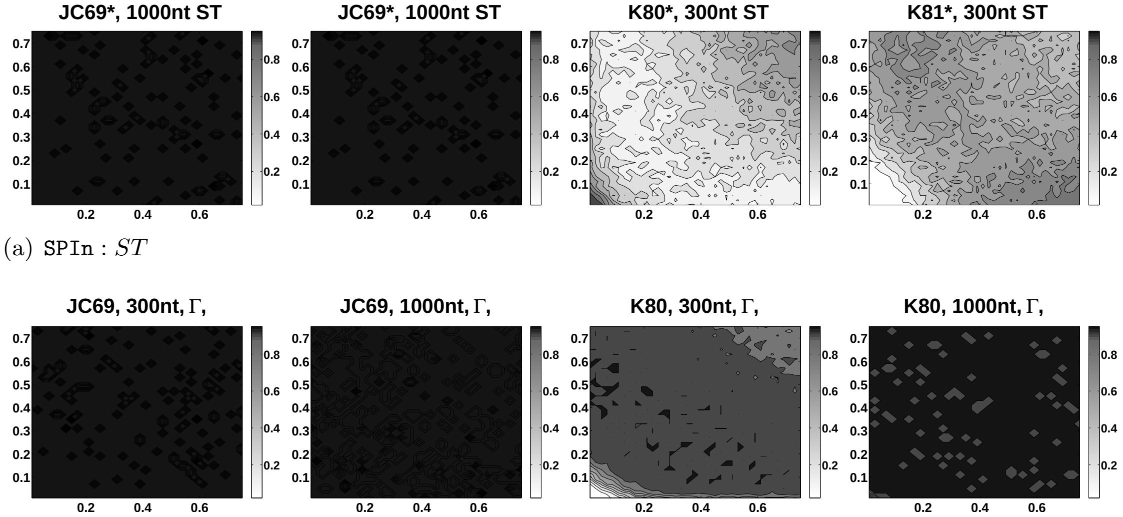
(b) *jModelTest*

(c) SPIn

(d) *jModelTest*

Figure 9.3: Plots of the fraction of correctly identified models for multiple sequence alignments of lengths 300 and 1,000 along 2-tree mixtures on quartet trees on the same tree topology (MST) under JC69* and K80*; SPIn: (a); *jModelTest*: (b); and on different tree topologies (MDT) under JC69* for 300nt and 3000nt; SPIn: (c); *jModelTest*: (d). The parameters vary in the quartet tree space: (a, b) of Huelsenbeck (1995). Fractions are displayed in grey-scale ranging from 0% in white to 100% in black. Corresponding average recovery rates are given in Table 9.1(c) and (d).

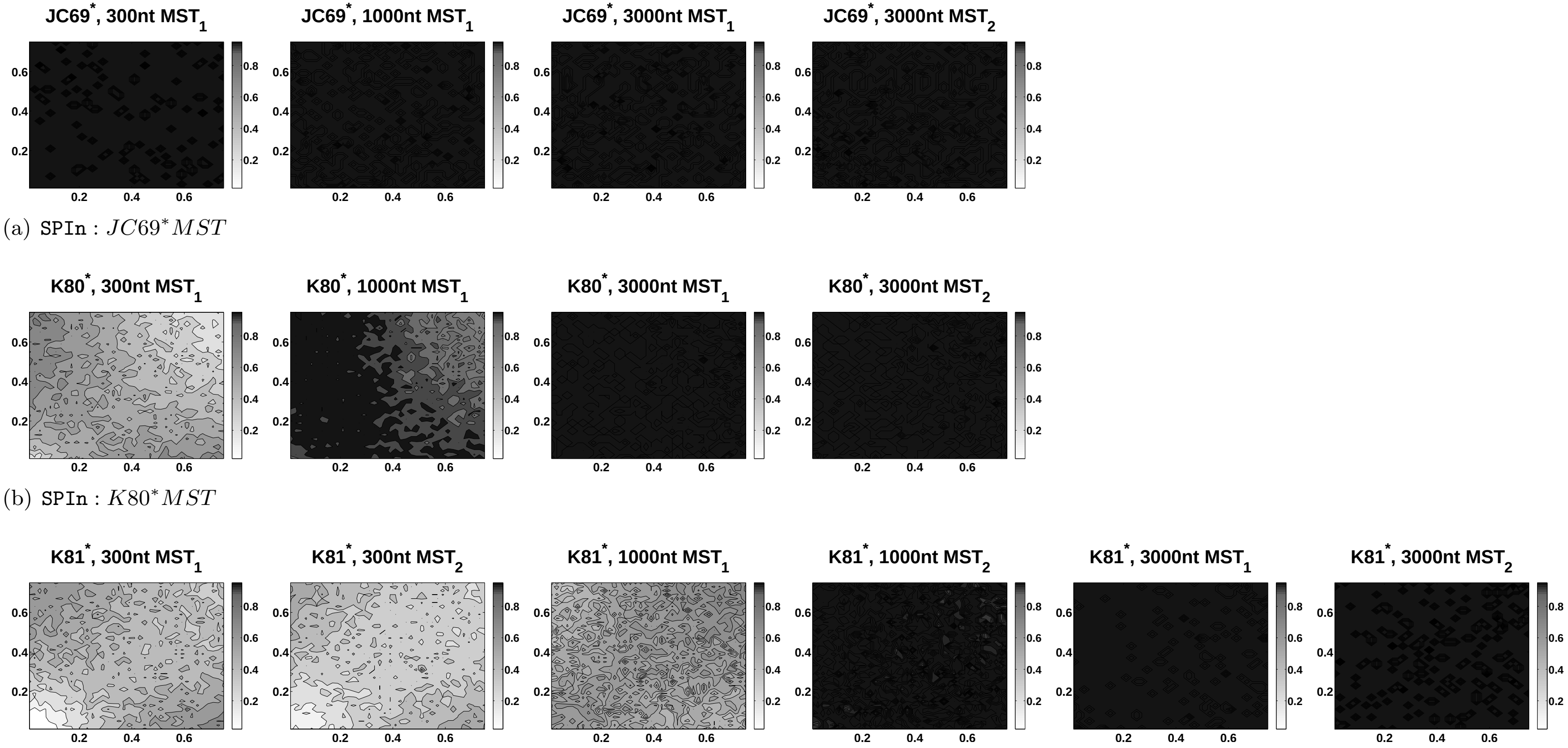
Figure 9.4: Performance assessement of SPIn.
Plots of the fraction of correctly identified models for multiple sequence alignments of varying lengths under discrete-time models on a single tree (a); and under continuous-time models with 2 Γ-rate classes (b). The parameters vary in the quartet tree space: (a, b) of Huelsenbeck. Fractions are displayed in grey-scale ranging from 0% in white to 100% in black.



(a) SPIn : ST

(b) SPIn : Γ

Figure 9.5: Performance assessement of `SPIn`. Plots of the fraction of correctly identified models for multiple sequence alignments of varying lengths under discrete-time models on quartet trees with the same tree topology (MST). The parameters vary in the quartet tree space: (a, b) of Huelsenbeck. Fractions are displayed in grey-scale ranging from 0% in white to 100% in black.

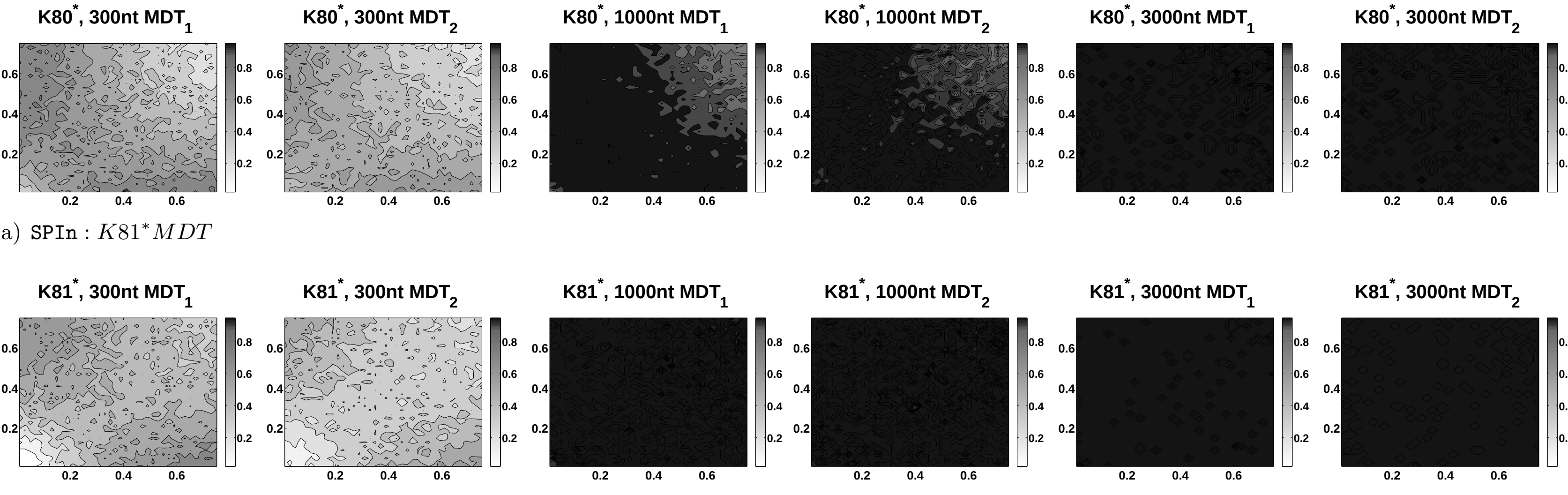


(a) `SPIn` : $JC69^{*}MST$

(b) `SPIn` : $K80^{*}MST$

(c) `SPIn` : $K81^{*}MST$

Figure 9.6: Performance assessement of `SPIn`.
Plots of the fraction of correctly identified models for multiple sequence alignments of varying lengths under discrete-time models on quartet trees with different tree topologies (MDT). The parameters vary in the quartet tree space: (a, b) of Huelsenbeck. Fractions are displayed in grey-scale ranging from 0% in white to 100% in black.



(a) `SPIn` : $K81^{*}MDT$

(b) `SPIn` : $K81^{*}MDT$

Larger trees on real-life topologies. In order to investigate the performance of `SPIn` when the number of OTUs is larger, we ran the tests on multiple sequence alignments simulated on two topologies inferred for real-life sets of species. As before, *evolver* package *(PAML)* was used to generate 100 multiple sequence alignments in the following settings: continuous-time `JC69` model with three discrete Γ-rate classes and length $5,000$ on the 9-taxon drosophila tree (Pollard et al., 2006a; Clark et al., 2007) and `HKY` (Hasegawa et al., 1985) model with four Γ-rate classes, transition/transversion ratio of $\kappa = 2$, nucleotide frequencies of $\pi_{\mathtt{A}} = \pi_{\mathtt{C}} = 0.1$, $\pi_{\mathtt{G}} = \pi_{\mathtt{T}} = 0.4$ and length $1,000$ along the 12-taxon T12b yeast tree (Marcet-Houben and Gabaldón, 2009); see Figures 1.5 and 9.3. In both cases the parameter α of the Γ distribution was set to 0.5.

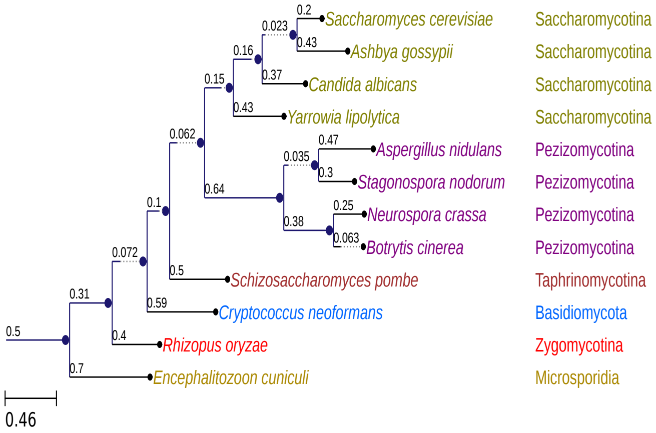


Figure 9.7: Phylogenetic tree used for simulations: 12-taxon fungal tree T12b (Marcet-Houben and Gabaldón, 2009).

Though the tree of drosophila has fewer sequences than the fungal tree, its branches are shorter, which in practice will lead to fewer different observed nucleotide patterns at the leaves. Therefore, in this case we simulated longer alignments of $5,000nt$. In both data sets `SPIn` recovered the model that the data was sampled from in 100% of the cases.

In addition, we tested the performance on the 10-taxon primate tree model obtained from Fujita et al. (2010) under continuous-time `JC69` and `K80` 3- and 4- tree mixture models. Since primate species are closely related, the resulting tree will have short length and presents challenges for model inference. We found that for 100% model recovery the required alignments lengths were on average $30,000$. Although this number might appear large, it is not unrealistic with the growing availability of complete genomes.

The method presented here is based on the nucleotide patterns recorded at the leaves of the tree, therefore it is better suited for more diverged trees. In practice, including distinct clades or an outgroup (as seen in the trees used here for simulations) will significantly improve the accuracy of model recovery.

Comparison to existing methods. Existing phylogenetic packages, as mentioned in the Introduction, rely on a similar model testing principle: an initially inferred phylogeny is used to select a model for subsequent tree inference. We decided to compare the performance of SPIn to that of *jModelTest*, which is a popular package designed specifically for model selection.

We are aware that *jModelTest* was not created to deal with the discrete-time mixture data. In order to allow maximum comparability between the two methods, we chose the following settings for the command line version of *jModelTest*: AIC_c criterion with the option of 5 models, enabled invariant sites and two gamma classes ($-AIC_c$ `-s 5 -i -g 2`). This ensured a fair comparison as the pool of models activated in *jModelTest* was contained within the models we considered. Although *jModelTest* supports neither discrete-time Markov models nor mixtures on a single or different tree topologies, we found it interesting to evaluate its performance on this type of data.

The results for the continuous-time JC69 and K80 models on a single tree are shown in Figure 9.2(b) and Table 9.1 (d). The average model recovery was 60% and did not depend on the length of the alignments. In comparison to the continuous-time models, the average recovery for the ST data under the discrete-time models dropped to 56% for the JC69* model, 37% for the K80* and 49% for the K81* models. Interestingly, the recovery rate was found to be worse with an increase of the alignment length from 300 to $1,000$, see Figure 9.2(d) and Table 9.1(d).

The same trend, though with a slightly lower impact, was found for 2-mixture data on the same topology, MST_1, under the K80* model: the mean recovery decreased from 41% in the $300nt$ data set to 37% for $1,000nt$ (Fig. 9.5(b)). The average detection for both MST_2 and MDT_2 data sets under JC69* dropped with an increase of the alignment length from 67% and 65% ($300nt$) to 56% and 45% ($1,000nt$), respectively (Tab. 9.1(d) and Fig. 9.5(b), 9.3(d)). The average model recovery on the MDT_1 data set was found to be the lowest (45%) among all the test for JC69* model.

Since SPIn was designed specifically to deal with phylogenetic mixtures and non-homogeneous data, the method outperforms *jModelTest* for the alignments generated under discrete-time models on single and mixture of trees. This result is due to the fact that, as proved in section 6.3, the linear invariants are strictly model specific and derived from the properties of the nucleotide substitution matrices as opposed to the exponential rate matrices.

In species tree reconstruction an assumption of a single tree topology is reasonable and the data is usually composed of the alignments of single copy homologous genes. However, though the tree topology remains the same, the branches might differ in lengths along the alignment, thus it becomes a mixture model. Unless the inference is performed on each block separately allowing for non-homogeneity of the rates at different lineages, this fact is not accounted for by the existing methods. In such instances, as shown in the above comparison, an incorrect model is very likely to be selected and this in turn may confound the tree inference. Though it was found that in some instances an approximated model might allow for recovering the species topology, the parameter estimates will not be correct. It can be seen in the results presented here that the methods accounting for mixtures increase the reliability of the results.

9.3 Discussion

SPIn uses linear invariants defining the spaces of all phylogenetic mixtures under a given model. The structure of a phylogenetic mixture model, for instance the number of components and tree topologies, is allowed to vary freely. While more statistical work is required to better address scenarios where a large number of sequences must be handled simultaneously, tests on simulated data coming from a single tree as well as mixtures of trees suggest that SPIn correctly identifies the underlying model in cases that proved difficult for existing methods.

Another issue regarding some of the existing methods is the tendency to select complex models. For instance, as found by Nguyen et al. (2011), in the analysis of 6,171 protein coding regions, the GTR class of models was selected in more than 70% of the cases (see Tab. 3 of Nguyen et al. (2011)). This was also the case for the protein-coding DNA alignment (PF02724) from the PANDIT database (Whelan et al. (2006)) analyzed by these authors. As shown in the quoted paper, the tree topology inferred under the $GTR + I + \Gamma$ (invariable sites and Gamma rates) model is incongruent with the accepted phylogeny However, using JC69 $+ I + \Gamma$, the tree topology is correctly recovered. We have analyzed this data set and the model selected by SPIn is in fact JC69*. This provides evidence that SPIn does not always choose most complex models for real data sets.

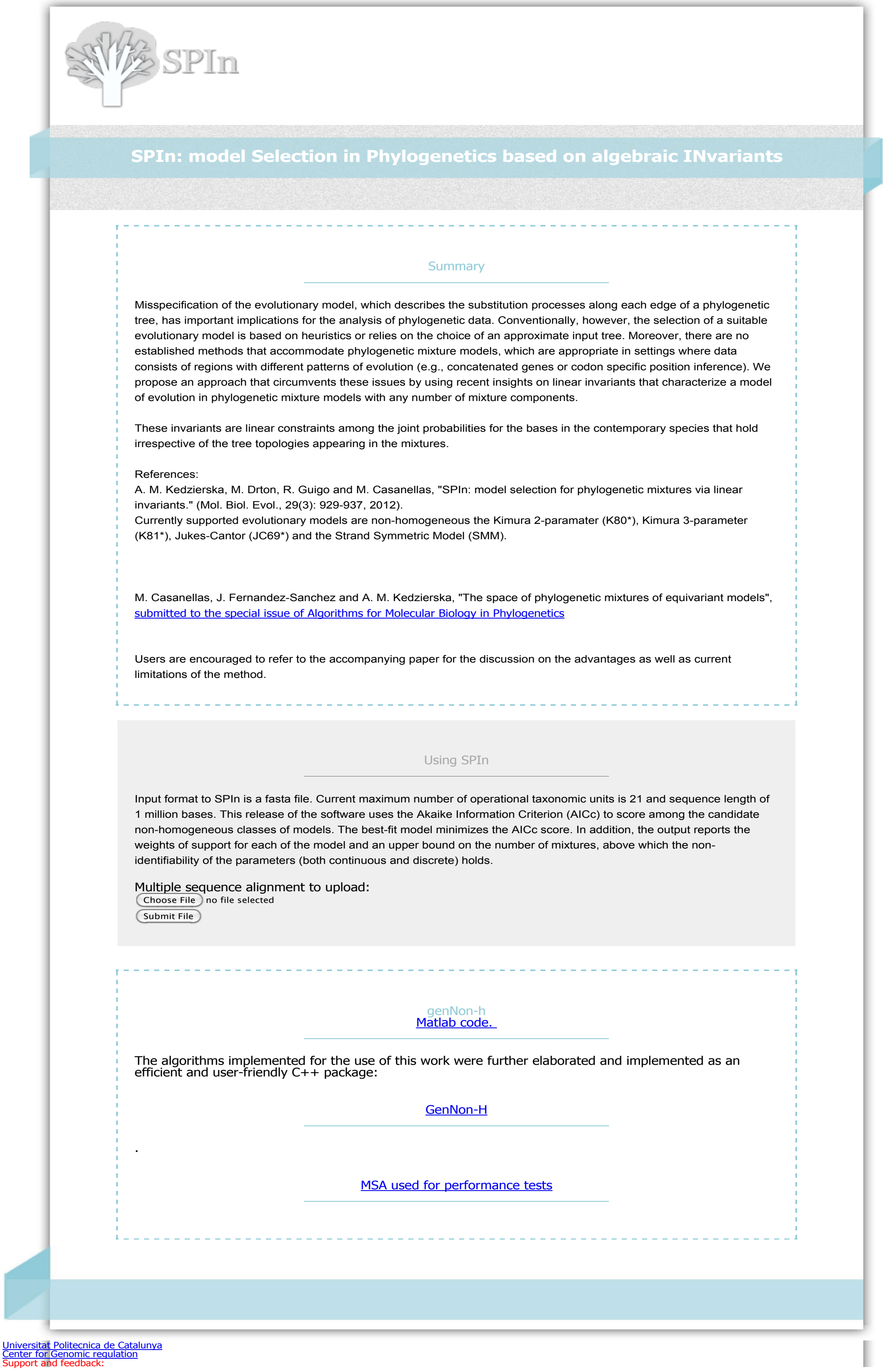
We propose using SPIn as a first inference step to discriminate between mixtures on the discrete-time models introduced in section 1.4: JC69*, K80*, K81*, SSM. If, for instance, the data supported JC69*, further analysis could address the question of whether an unmixed JC69, JC69+Γ, or JC69+Γ+I fits the data better. One could also investigate the number of different tree topologies that should be taken into account.

In the current version of the program gaps and ubiquitous characters are removed from the alignment. Note that the number of invariants for each model is 4^n minus its dimension. Although this number is exponential in n the implementation of SPIn uses only the invariants containing the patterns observed in the data. As the length of the alignment is not exponential in n, the algorithm in fact uses a subset of invariants. This approach significantly speeds up the algorithm. Current implementation limits the maximum number of input species in SPIn to 21. However, an ongoing work is to extend this number to increase applicability to the modern real-life analyses.

Here we demonstrated good performance for up to 10 species with up to 100,000 sites when using AIC_c. Another option is Bayesian Information Criterion (Schwarz, 1978; Burnham, 2004), however, our experience showed that the large sample properties of the BIC are reasonable for short alignments and sparse data. Ongoing work on sampling based statistical inference aims at extending the applicability of SPIn to larger number of species. This said, the patterns and rates of evolution which characterize functional elements depend on their location within the genome, the $G + C$ content of the region, synonymous codon site selection (features addressed by accounting for mixture models) and tend to be clade-specific (Pollard et al., 2010). In large studies, we recommend grouping the sequences and performing the selection on such subsets. Also, in order to deal with incomplete or new genomes, future release of `SPIn` will include methods to deal with highly sparse data and short alignments.

An attractive feature of `SPIn` is its speed. Irrespective of the model considered, the time to run `SPIn` on a 2-core Intel machine (2.40GHz) with 48 GB of RAM on a multiple sequence alignment of 4 OTUs of length 300 was on average 0.014s, 0.020s for length 3000 and $0.177s$ for 10-taxon multiple sequence alignments of length $30000nt$. As a comparison, in the latter case *jModelTest* took $6m28s$.

There is a number of improvements to the method. We believe that more information can be extracted using linear invariants, e.g. composition of topologies in the mixture, expanding the spectrum of available models. We discuss this ongoing and future goals in section 11.



SPIn: model Selection in Phylogenetics based on algebraic INvariants

Summary

Misspecification of the evolutionary model, which describes the substitution processes along each edge of a phylogenetic tree, has important implications for the analysis of phylogenetic data. Conventionally, however, the selection of a suitable evolutionary model is based on heuristics or relies on the choice of an approximate input tree. Moreover, there are no established methods that accommodate phylogenetic mixture models, which are appropriate in settings where data consists of regions with different patterns of evolution (e.g., concatenated genes or codon specific position inference). We propose an approach that circumvents these issues by using recent insights on linear invariants that characterize a model of evolution in phylogenetic mixture models with any number of mixture components.

These invariants are linear constraints among the joint probabilities for the bases in the contemporary species that hold irrespective of the tree topologies appearing in the mixtures.

References:
A. M. Kedzierska, M. Drton, R. Guigo and M. Casanellas, "SPIn: model selection for phylogenetic mixtures via linear invariants." (Mol. Biol. Evol., 29(3): 929-937, 2012).
Currently supported evolutionary models are non-homogeneous the Kimura 2-paramater (K80*), Kimura 3-parameter (K81*), Jukes-Cantor (JC69*) and the Strand Symmetric Model (SMM).

M. Casanellas, J. Fernandez-Sanchez and A. M. Kedzierska, "The space of phylogenetic mixtures of equivariant models", submitted to the special issue of Algorithms for Molecular Biology in Phylogenetics

Users are encouraged to refer to the accompanying paper for the discussion on the advantages as well as current limitations of the method.

Using SPIn

Input format to SPIn is a fasta file. Current maximum number of operational taxonomic units is 21 and sequence length of 1 million bases. This release of the software uses the Akaike Information Criterion (AICc) to score among the candidate non-homogeneous classes of models. The best-fit model minimizes the AICc score. In addition, the output reports the weights of support for each of the model and an upper bound on the number of mixtures, above which the non-identifiability of the parameters (both continuous and discrete) holds.

Multiple sequence alignment to upload:

genNon-h
Matlab code.

The algorithms implemented for the use of this work were further elaborated and implemented as an efficient and user-friendly C++ package:

GenNon-H

.

MSA used for performance tests

Universitat Politecnica de Catalunya
Center for Genomic regulation
Support and feedback:

Chapter 10

Conservation patterns in biotypes of the *GENCODE* annotation.

In this chapter we present the results on the study of evolutionary patterns in different genetic biotypes using methodologies developed in the previous chapters. The analysis was performed on the version 3c of the *GENCODE* human gene and transcript annotation (UCSC Genome Browser, Fujita et al. (2010); Harrow et al. (2006)) on the hg19 version of the human genome ($gencode.v3c.annotation.GRCh37$, see section 1.2).

The human genome was partitioned in segments according to the GENCODE biotype of the annotated transcripts and subsequently to type of the genetic elements within. Transcripts in *GENCODE* are assigned a biotype, which reflects their biological functions. In this study, we used the following: BIOTYPE={protein, lncRNA, pseudogene, protein/lncRNA, protein/pseudogene, lncRNA/pseudogene}
and ELEMENT={exon, intron, UTR, CDS, mix}. Here, "protein" refers to the protein coding transcripts, "lncRNA" to the long non-coding RNAs, "mix" is intron and exon, and "UTR" includes both 3' and 5' untranslated regions. In this analysis we merged the annotation of the processed and nonprocessed pseudogenes into one single pseudogene biotype or functional class (see Aheng et al. (2007) and the references within).

We have, therefore, obtained a partition of the human genome sequence into non overlapping segments (a segmentation) in which the segments correspond to one of 18 functional classes (see below) defined as BIOTYPE.ELEMENT (note: that not all combinations of biotype and element are valid).

Having built this partition, we investigated the following questions:

- Do different partitions show distinct patterns of conservation? Is there a clear support for a particular model?
- Are the patterns for pure (e.g. exonic lncRNA), mixed (e.g. mixed lncRNA) or multi-label (exonic lncRNA/pseudogenes) classes different?
- Are these evolutionary patterns reflected in the estimates of the branches in the species tree?
- Is there a specific model that best characterizes conserved regions? On the other hand, is there a best-fit model for neutrally evolving regions?
- Does model information allow for more accurate estimation of the branch lengths in the phylogenetic tree?

The last question is an interesting and disputable one. While ad-hoc model choice is an advantage, it is a common practice to perform phylogenetic analyses choosing a model heuristically (see Chap.9) . It is believed that more complex models are a better choice.

However, as discussed in section 9.3 for the real-life protein-coding data set, it is in fact the `JC69`* model (as opposed to the $GTR+I+\Gamma$ model) that supports the correct topology (Whelan et al., 2006).

The above annotation-induced partition of the genome, we obtained genome segments showing different levels of inter-species conservation that allowed us to investigate the above questions in detail.

10.1 Conservation vs functionality

At present, the proportion of the human genome that encodes functional elements is unknown. Both coding and non-coding regions are affected by negative selection (Shabalina and Kondrashov, 1999; Makalowski and Boguski, 1998). Comparative genomic of human, dog, mouse and rat by Kamal et al. (2006) revealed that about $5-6\%$ of the human genome is thought to be under purifying selection, of which strinkingly only $1-2\%$ lies within the protein-coding sequences. The remaining parts are conserved non-coding elements. As reported in (Siepel et al., 2005), over 32% of the highly conserved sequences lie within the unannotated regions. Recently, TheENCODEProjectConsortium (2011) (cf. references within) estimated the percentage of base pairs of the human genome under purifying (or negative) selection to be between 3%–8%. The authors suggest this number to be underestimated due to the faults of current phylogenetic methodology. It was suggested by Pheasant and Mattick (2007) that the bound of only 5% of the genome coding for functional information should be increased. Failure to detect functional elements that are short or fragmented is a serious drawback to the methods. At the same time, it argues in favour of using phylogenetic mixtures. This was confirmed in the case study of splicing regulators in section 2.2, where synonymous positions were shown to be under negative selection not directly related to the protein coding potential.

Let us first look at the biological types that we have chosen to investigate here, and their characterization in terms of conservation.

Protein-coding genes are the regions of the genome for which the RNA transcript is subsequently translated into protein. The transcribed part of a gene is composed of introns and exons (see Chap. 1). Introns are the non-coding sequences, which are removed from the primary transcripts. The number and size of introns varies between the organisms. In vertebrates introns constitute the major part of the protein-coding genes, some being thousands of nucleotide in length. Human introns can reach even greater lengths. In general, introns are not expected to be conserved across the species, however, they may contain stretches of conserved regions (Sugnet et al., 2006). Others factors playing key role in intron conservation can be, among others, the presence of the stem-loop structures and overlapping transcripts (Barrette et al., 2001).

Due to their importance in protein synthesis, exons are expected to be the most conserved among the functional classes considered here. The level of conservation varies and depends on a variety of factors, e.g. the number of splice variants a given exon belongs to and its functionality, the type and conservation of the splice sites and adjacent intron lengths (see e.g. Irimia et al. (2008)).

Untranslated regions (UTRs) of the protein-coding genes are in general less conserved than the protein-coding regions, however, both $3'$ and $5'$ UTRs contain regulatory sequences (Churbanov et al., 2005; Wegrzyn et al., 2008; Chen and Rajewsky, 2006). Thus, sequence conservation of the UTRs can be significant and, in some cases, even higher than the neighbouring CDSs (i.e. Spicher et al. (1998)). For instance, in mammals the conservation of the UTRs was found to be positively correlated to their base composition (Shabalina et al., 2003).

Long non-coding RNAs (lncRNA) are defined as non-protein coding transcripts “longer than 200nt”. Although they show a general low across-species conservation, which by some authors is interpreted as potential lack of functionality (Struhl, 2007; Marques and Ponting, 2009), selection may act on small regions in the long lncRNA transcript. Many lncRNAs contain elements that are under purifying selection or lie within the regions conserved due to their function (i.e. lie in the promoter regions or are functional in splicing; Ponjavic et al. (2007); Pollard et al. (2006b)).

Pseudogenes usually originate from duplication of functional genes (Zhang and Gerstein, 2004), but have subsequently loss functionality. Regions annotated as pseudogenes are oftentimes used to model neutral evolution. The extent of conservation of pseudogenes seems to be disputable. In opposition to the expectation, pseudogenes are not free of negative selection and were found to be conserved across species and functionally active. In Balakirev and Ayala (2004, 2003), the authors support the hypothesis that pseudogenes are to be considered as protogenes, which are the DNA sequences with the potential of becoming new genes. On the other hand, characterization of pseudogenes within *ENCODE* Aheng et al. (2007) showed that most pseudogenes evolve neutrally.

Ancient repeats. As a control set of the background rate of neutral evolution we chose ancient repeats (ARs) from Ensembl (obtained via Repeat Masker). With some exceptions to the rule, ARs are largely nonfunctional.

The 18 resulting functional classes (17 biotype.element + AR) are shown in the first column of Table 10.1. For example, *lncRNA/pseudogene.mix* corresponds to the genomic regions annotated as exon and intron (i.e. of different isoforms) of a lncRNA and a pseudogene, while regions under the label *protein.intron* were annotated as protein coding introns only.

Table 10.1: Partitions of the annotation into different biological domains.

partition	length (bp)	model
lncRNA.exon	810.531	SSM
lncRNA.intron	20.857.941	SSM
lncRNA.mix	318.361	SSM
lncRNA/pseudogene/exon	52.610	K80*
lncRNA/pseudogene.mix	52.466	K80*
lncRNA/pseudgene/intron	288.186	SSM
pseudogene.exon	328.210	SSM
pseudogene.intron	993.364	SSM
pseudogene.mix	7.819	K80*
protein.intron	61.966.684	SSM
protein.CDS	7.467.354	SSM
protein.mix	652,091	SSM
protein.UTR	9.725.790	SSM
protein/lncRNA.intron	1.5804.897	SSM
protein/lncRNA.mix	1.794.411	SSM
protein/pseudogene.intron	361.187	SSM
protein/pseudogene.mix	85.971	K80*
ancient repeats	169.449	SSM

10.2 Pipeline

The analysis proceeded in the following 5 steps:

Step 1: Data extraction. From the *ENCODE* whole genome multiple sequence alignments (MSAs) of 46 species we selected the MSAs of 6 taxa: human, macaque, mouse, rat, cow and dog (hg19, canFam2, bosTau4, mm9, rn4, rheMac2, cf. Section 2.1).From the genome partition on functional classes, we extracted the MSAs corresponding to each segments. Columns containing gaps or ambiguous characters were removed from the set. Table 10.1 shows the total sizes of these data sets.

Step 2: Choosing the best-fit model using SPIn. Assuming nonhomogeneity (and no tree topology) we used SPIn to choose the best fit model for the data sets extracted in Step 1. We ran the analysis “globally” by running SPIn on the merged alignments from segments for each functional class, and “locally” by looking at each extracted segment separately (see next).

Step 3: Global comparison. We inferred the model for the merged alignments of all segments within each functional class (biotype.element).

Step 4: Local comparison. We compared the distribution of the supported models within each partition (biotype.element) separately. We performed model selection on each of the extracted segments provided they were longer than 100nt.

Step 5: Use Empar to infer branch lengths of the species tree. Under the discrete-time model inferred for the concatenated sequences (in Step 3) and on the species tree, we estimated the continuous parameters of the model and calculated the branch lengths for each partitions.

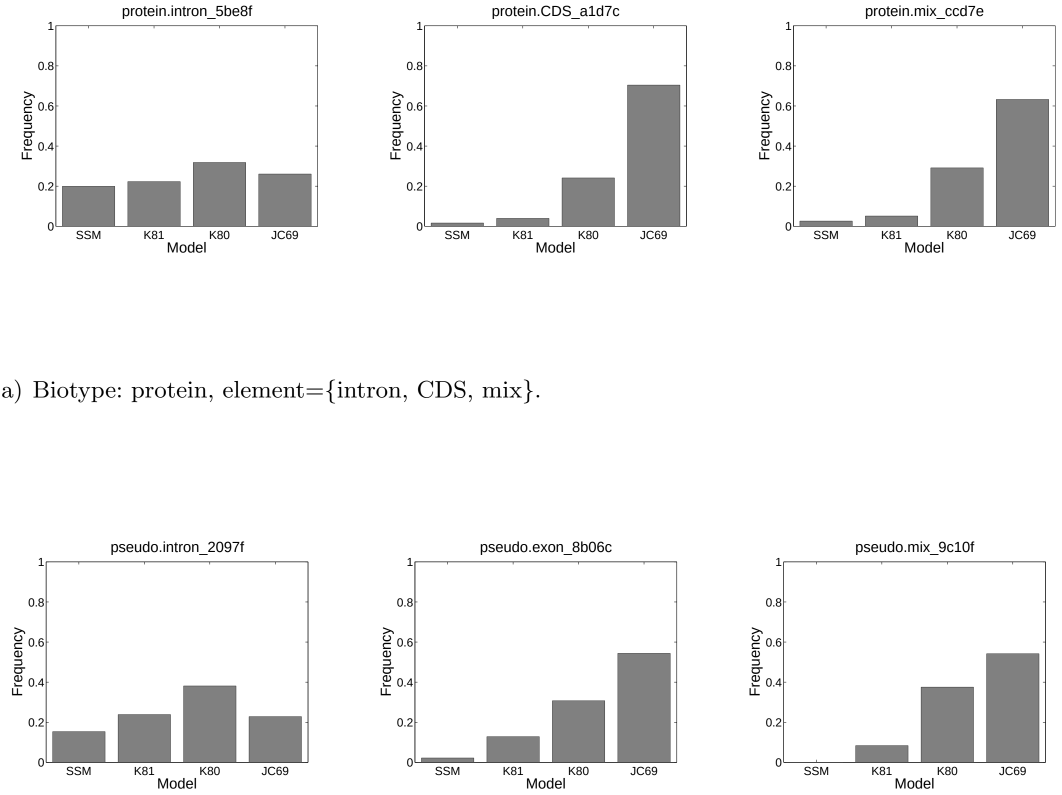
In Step 2 we consider the equivariant models (see Sec.5.5): JC69*, K80*, K81* and SSM as currently implemented in SPIn. Though the pool of models is limited, these models are nonhomogeneous models and allow different rates in distinct lineages, e.g. SSM is a nonhomogeneous version of the HKY model Hasegawa et al. (1985)). In addition, a model chosen by running SPIn is valid for any phylogenetic mixture under this model. Therefore, the models are in fact a much broader class. For example, the model given by Jin and Nei (1990) is a continuous-time Kimura80 (Kimura, 1980) model with the discrete Gamma rates under the assumption that the data evolved on a single tree topology. On the other hand, the K80* model allows both the Gamma rates and mixtures on different trees. Model choice must be an optimal trade-off between the complexity and the amount of data– more data allows to include more parameters. For example, K80* has 2 parameters, while the SSM model has 8 free parameters, 6 per edge in the substitution matrices and 2 in the root distribution. Overparameterization of the models may lead to non-identifiability, which means that the parameters cannot be recovered from the observed data.

10.3 Results and discussion

Only two models were supported in the global analysis. From Table 10.1 we note that SSM is preferable in larger data sets, while the K80* was selected in 4 cases for shorter data. This is in agreement with the intuition that long strands of DNA will support a more flexible model reflecting double-strandedness of the DNA. Sufficiently large amount of data (large alignments) allows for viable estimation of the parameters, justifying the use of the more complex models.

Models selected for individual elements in each partition (i.e. local analysis) were plotted as normalized histograms, i.e. relative frequencies of the models for all MSAs of a given partition (see Figs.10.1, 10.2 and 10.3).

Biotype: protein, element={intron, CDS, mix}. It can be seen that the intronic sequences show some support towards K80* and are uniform over the remaining models (see Fig. 10.1(a)). In the CDS regions, which are expected to be most conserved among the classes, we observed a significant support towards the JC69* and lack of it towards the SSM and the K80* models. In the regions overlapping the CDS and introns of different transcripts, we observe a slightly weaker, but clear support towards the JC69* model. This suggests that the CDS signal is stronger and sequence conservation is comparable to that of the pure CDSs.



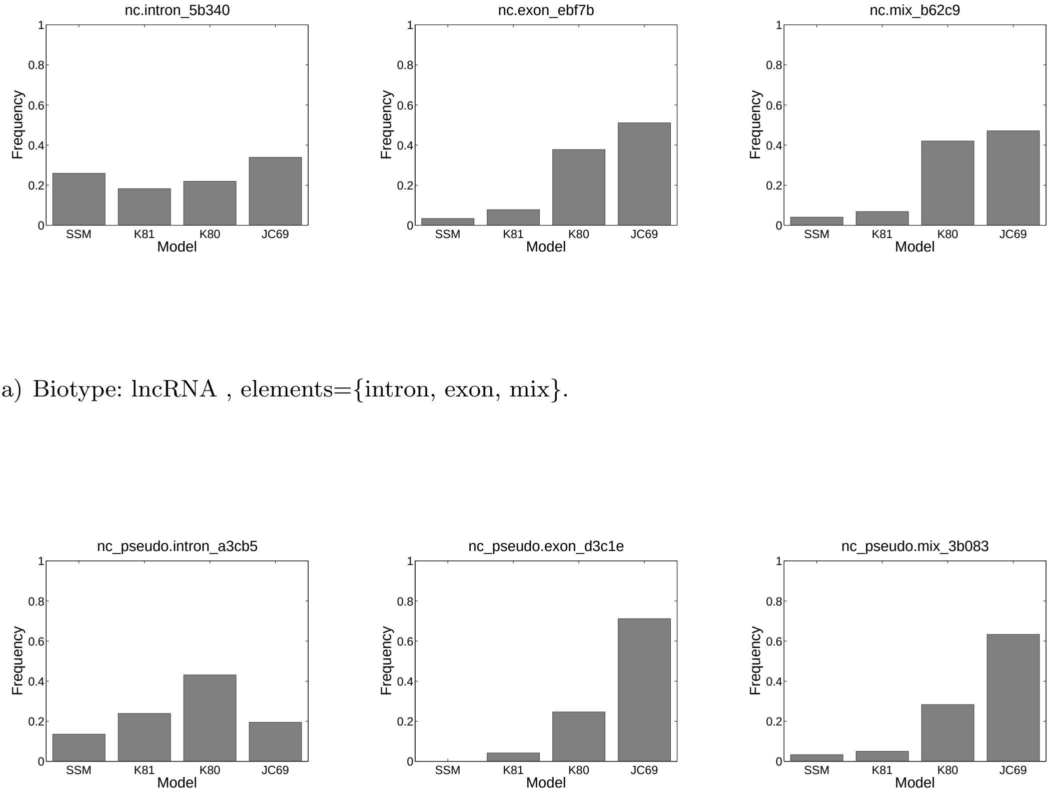
(a) Biotype: protein, element={intron, CDS, mix}.

(b) Biotype: pseudogene, element={intron, exon, mix}.

Figure 10.1: Histograms of the models inferred for the *GENCODE* biotypes of protein and pseudogenes.

Biotype: pseudogenes, element={exon, intron, mix}. From Figure 10.1(b) we see that the results for *pseudogene* biotype are comparable to those of protein-coding. The exonic and mixed regions show less significant preference towards `JC69`*with more visible support given to the `K80`* and `K81`* model.

Biotype={nRNA and lncRNA/pseudogenes}, element={intron, exon, mix}. MSAs annotated as intronic lncRNA show different pattern of distribution to the previosuly analyzed data sets (see Fig. 10.2(a)). They are uniformly distributed giving slightly more weight to the `SSM` and the `JC69`* models. Exonic and mixed regions resemble their correspondents in pseudogenes, with more support, however, towards the `K81`* model. This is further stressed in the regions annotated as both these classes: *lncRNA/pseudogene*. In fact, as seen in 10.2(b), exonic and mixed data sets of this biotype show high resemblance to the corresponding types of the *protein*-coding data (cf. 10.1(a)).



(a) Biotype: lncRNA , elements={intron, exon, mix}.

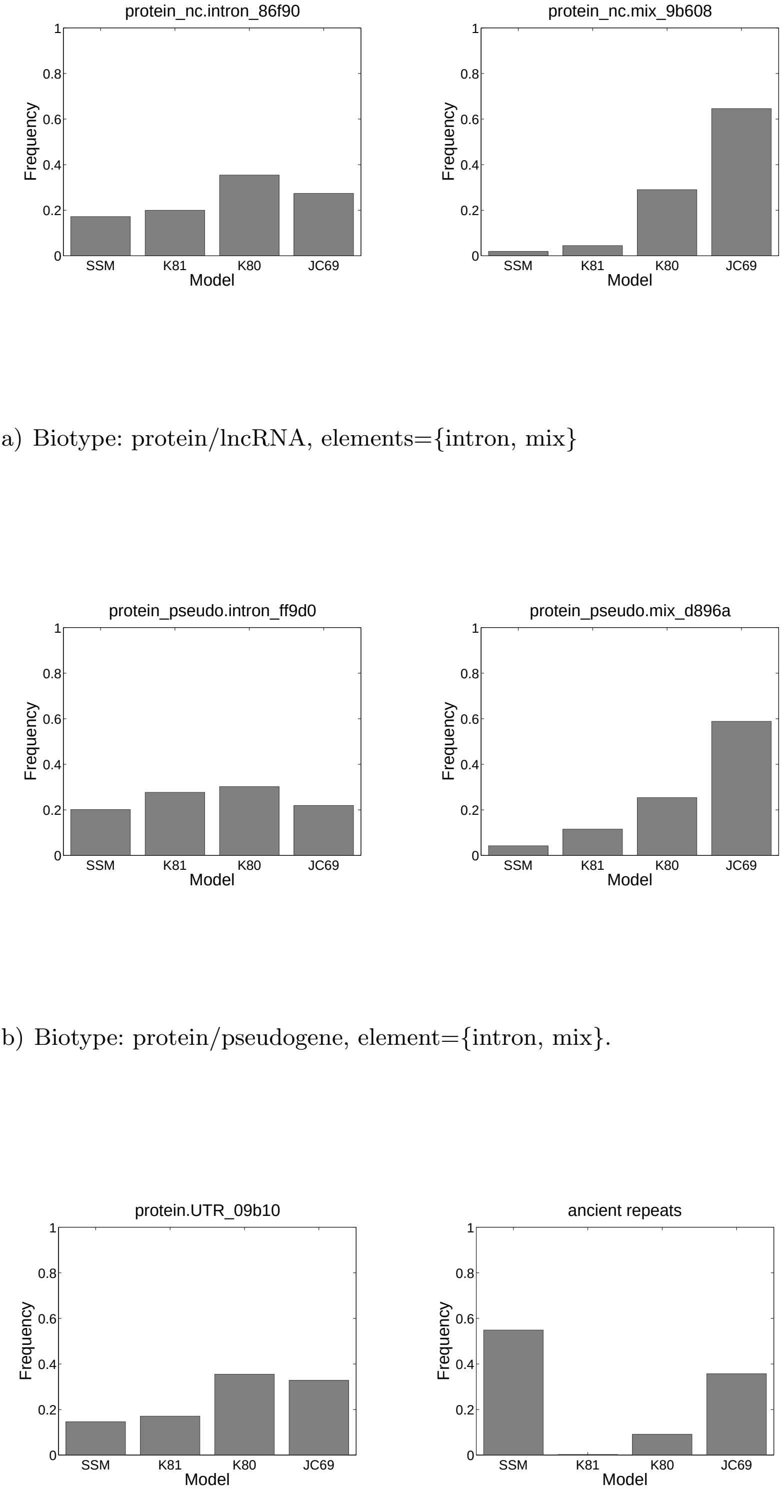
(b) Biotype: lncRNA/pseudogenes, elements={intron, exon, mix}.

Figure 10.2: Histograms of the models inferred for the *GENCODE* biotypes of *lncRNA* and *lncRNA/pseudogene*.

Biotype: protein/lncRNA, element={intron, mix}. Figure 10.3(a) shows the results for the regions annotated both as protein-coding and lncRNA. It can be seen that the intronic regions follow on the intronic patterns of all but the *lncRNA* data sets, showing high resemblance to the protein-coding introns. Similarly, mixed regions show great similarity to the protein-coding mixed regions (cf. Fig.10.1(a)). Thus, lncRNA signal seems to be playing a secondary role.

Biotype: {protein/pseudogenes}, element={intron, mix}. Intronic regions of the *protein/pseudogene* biotype are “an average” of the corresponding intronic distributions of the pure classes. Most visible support is given to the Kimura class of models: `K80`* and `K81`* (see Fig. 10.3(b)). This suggests that the sequence information in these regions might show distinguished differences in the transition/transversion ratio.

Again, in the exonic sequences the most supported evolutionary model is `JC69`*, and the resemblance to the patterns in the corresponding *protein* types is more prominent than in the *pseudogene* biotype alone.



(a) Biotype: protein/lncRNA, elements={intron, mix}

(b) Biotype: protein/pseudogene, element={intron, mix}.

(c) Biotype: protein, element: UTR; ancient repeats

Figure 10.3: Histograms of the models inferred for the *GENCODE* biotypes of *protein/pseudogene* and *protein/lncRNA*.

Biotype: protein/UTR and Ancient Repeats. These two sets are expected to have the least degree of conservation among all the data sets considered here. As depicted in Figure 10.3(c) we observe that the intronic sequences show similar patterns to the introns of the *protein/lncRNA* and *protein* biotypes with slightly more weight given to the K80* model. ARs differ significantly from the other distributions. Histograms plotted for the models selected in the regions covered by the ARs show the strongest support towards the SSM class among all the classes with some support given to JC69* (see Fig. 10.3(c)). From the results discussed thus far, this might suggest that a large portion of the ARs lies in the highly conserved regions (the JC69* fraction). On the other hand, we might hypothesize that the truly neutrally evolving sequences of the ARs fall into the SSM portion of the histogram.

10.3.1 Estimate of the branches in the species tree

We next used the species tree and the package Empar (see Chap. 8) to estimate the continuous parameters and the branch lengths under the model selected in the previous section.

In Figure 10.4 we plotted the results for *protein* and *lncRNA/pseudogene* biotypes, Figure 10.5 depicts the results for *protein/pseudogenes* and *lncRNA*, and Figure 10.6 for *protein/lncRNA*, *pseudogenes* and AR. As observed, the branches of the tree within all intronic regions are longer than those (even partially) annotated as exonic. For the CDSs the tree is the smallest with all its branches being the shortest among all the trees. Similar results were obtained for the exons of the *lncRNA/pseudogene* and *protein/lncRNA* biotypes. The trees of the exons and mixed *lncRNA* are very similar in lengths of the branches. Comparable to those are the branches estimated for the species tree in ARs, suggesting that in this set the corresponding regions of the MSAs might not evolve neutrally. As already mentioned above, many of these sequences seem to fall within the regions conserved across the genomes considered here.

By far the largest tree corresponds to the intronic *lncRNA*. The trees for other intronic regions are slightly shorter, but to a large extent comparable.

Summary. The key observation based on the results presented in this chapter is that the type of elments (exon vs intron) dominates the preferred choice of the model over the biotype. Thus, exons, irrespectively of whether they come from proteins, lncRNAs or pseudogenes, for the most part follow the JC69* model, i.e. the simplest of the evolutionary models. In contrast, introns, again whether from protein coding genes, lncRNAs or pseudogenes, show preference towards K80*. This is to some extent surprising, since one would have expected lncRNA (and most pseudogene) exons to evolve in a similar manner to that of introns and UTRs. Finally, ancient repeats, show a clearly differential pattern of evolution, following majoritarily SSM, which is the most complex of the models.

In addition, we make further observations:

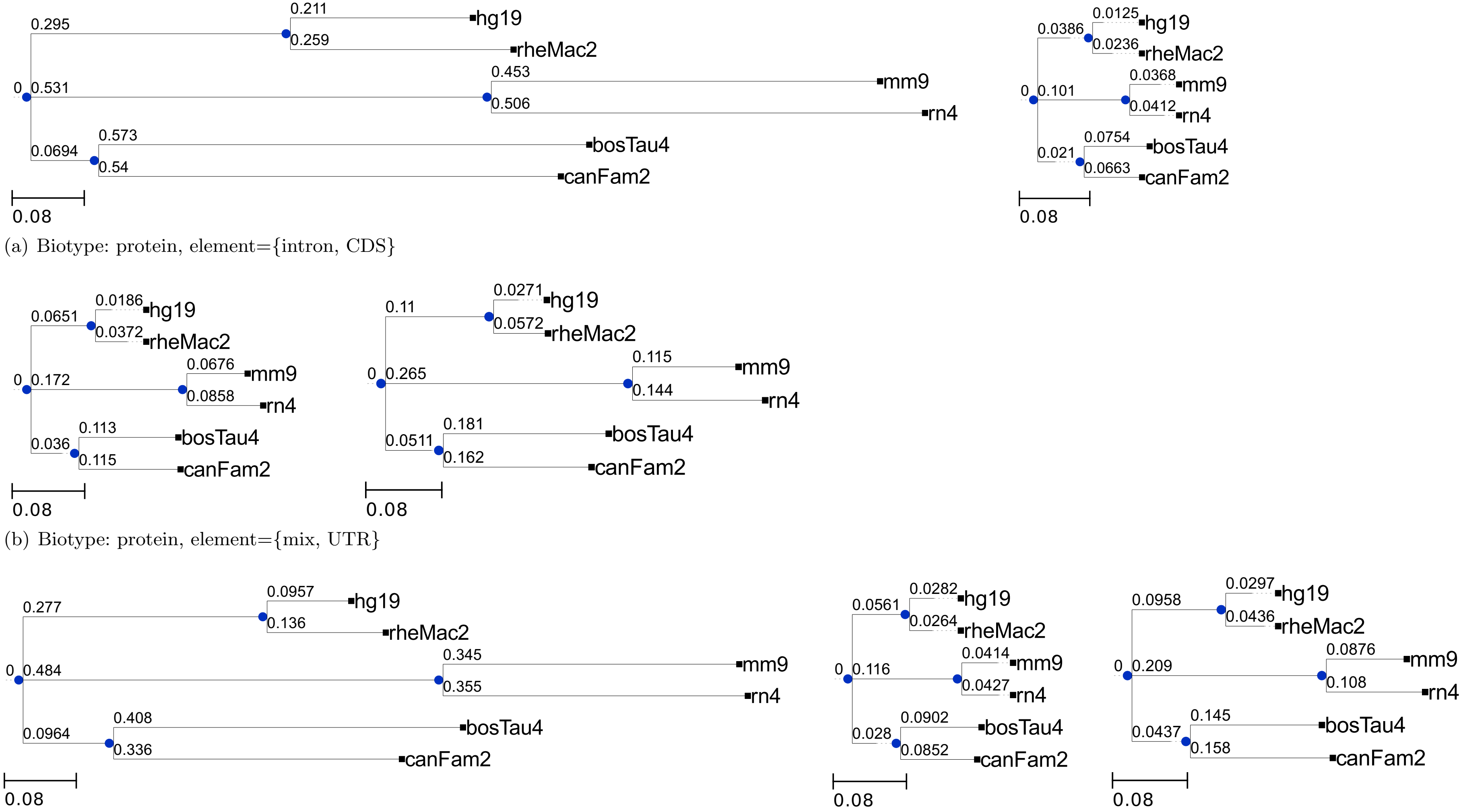
- The set of the models supported within each partition reveals similarities in the across-species conservation between the data sets. The distribution of the nonhomogeneous evolutionary models and the estimated lengths of the branches gave consistent results: the more uniform the distribution, the longer the branches estimated for the species tree. In general, the data sets with similar shapes of the distributions were found to have comparable trees, e.g. *protein/pseudogene* and *protein/lncRNA* of mixed type. The distributions with the highest support given to `JC69`* and `K80`* were found to have the shortest trees. We conclude that by looking at the models supported in a particular data set we can gauge its degree of evolutionary conservation.

- As expected, `SSM` is best suited for long alignments. However, its applicability is not limited to large data sets. In the analysis performed locally, we observed that the `SSM` model was absent in the regions expected to be conserved (i.e. exonic) with increasing support given to it in the intronic regions and a significant support in the ARs set. This suggests that `SSM` might be preferable in the neutrally evolving regions.

- The fairly simple `JC69`* model seems to be well suited for the conserved regions. It was selected in a large portion of the extracted alignments both in the protein-coding CDS and other exonic regions. In turn, using it to estimate the parameters gave rise to the trees with short branch lengths.

- The intronic type in the *lncRNA* and *lncRNA/pseudogene* biotypes show different pattern of model support. Exons and mixed types of *lncRNA/pseudogene* resemble more the respective sets in the data annotated as *protein*−coding. This might suggest that the pseudogenes are to some extent under purifying selection.

- Overall, ancient repeats might not be the best choice for a background neutrally evolving model. In comparison with the trees in the remaining partitions, the external branches of its estimated tree were short, i.e. the clade $(hg19, rheMac2)$ was comparable to that of exonic sequences. By and large, this finding suggests that a large part of the regions annotated as ancient repeats overlaps regions that are under negative selection. On the other hand, judging by the shape of the model distribution and the length of the estimated tree, intronic *lncRNA* set might be under strong positive selection.

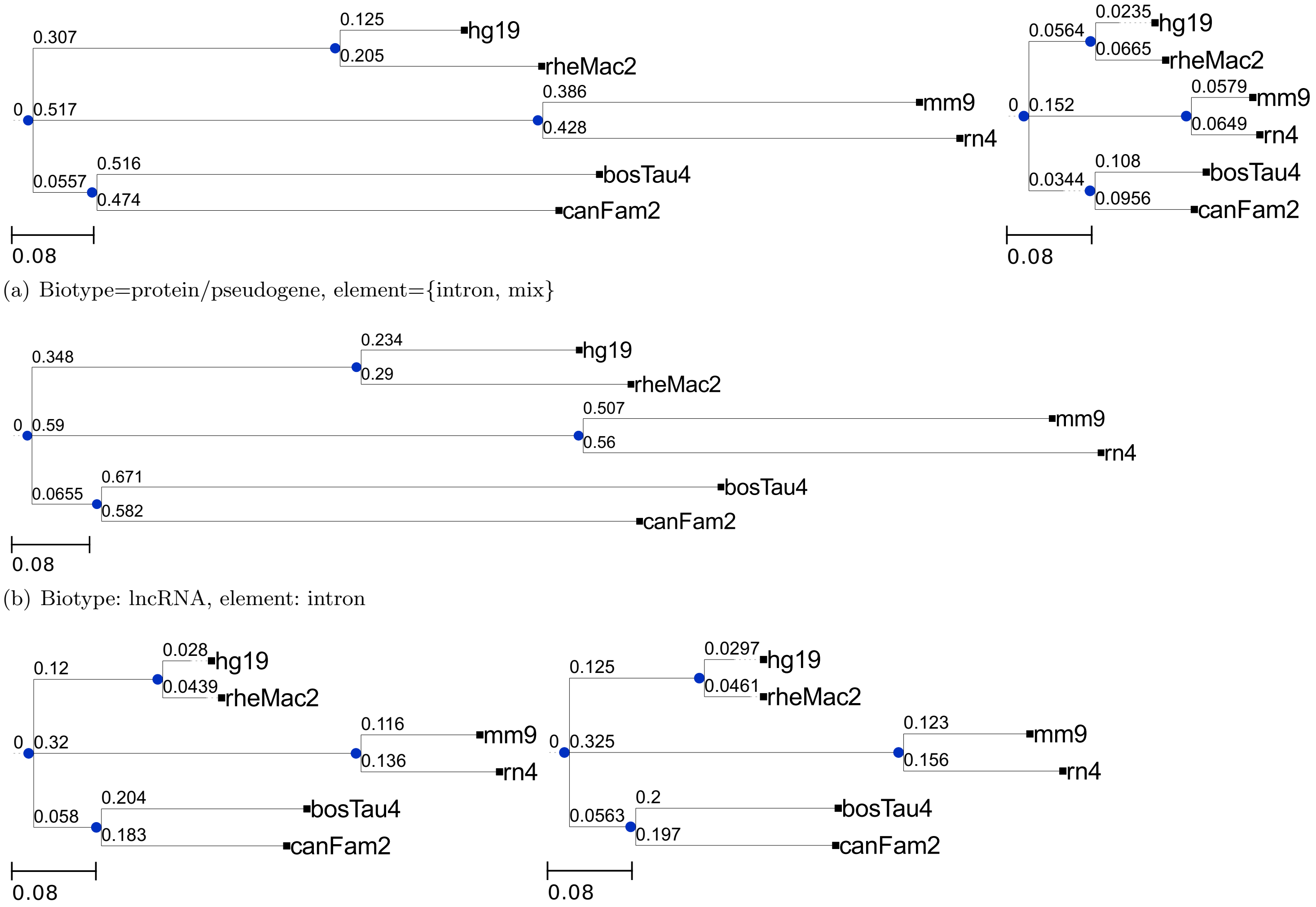


(a) Biotype: protein, element={intron, CDS}

(b) Biotype: protein, element={mix, UTR}

(c) Biotype: lncRNA/pseudogene, element={intron, exon, mix}

Figure 10.4: Phylogenetic trees for the *GENCODE* partitions (*labeled from left to right*, drawn using Marcet-Houben and Gabaldón (2009)).

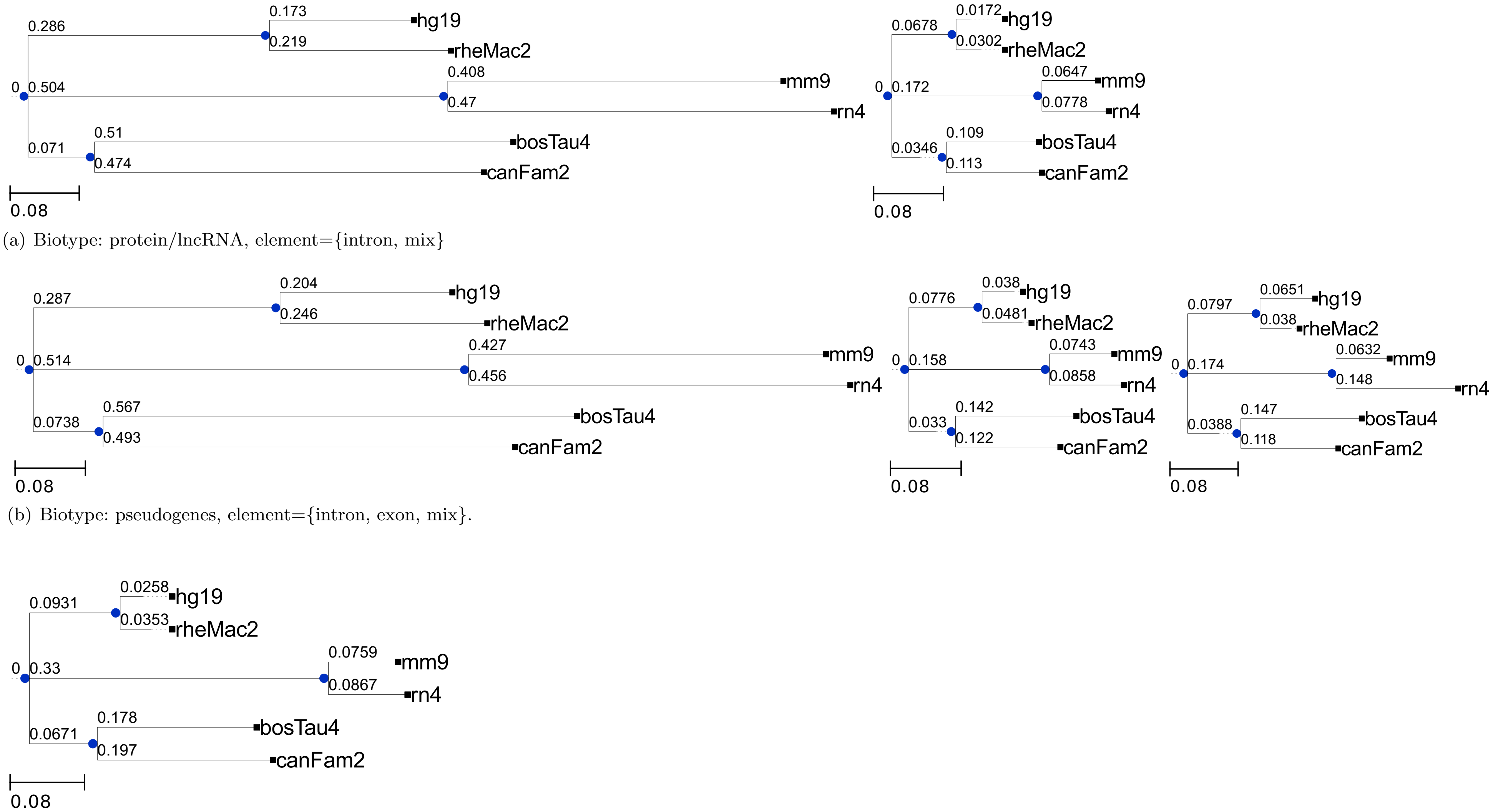


(a) Biotype=protein/pseudogene, element={intron, mix}

(b) Biotype: lncRNA, element: intron

(c) Biotype: lncRNA, element={exon, mix}

Figure 10.5: Phylogenetic trees for the *GENCODE* partitions (*labeled from left to right*, drawn using Marcet-Houben and Gabaldón (2009)).



(a) Biotype: protein/lncRNA, element={intron, mix}

(b) Biotype: pseudogenes, element={intron, exon, mix}.

(c) Ancient repeats

Figure 10.6: Phylogenetic trees for the *GENCODE* partitions (*labeled from left to right*, drawn using Marcet-Houben and Gabaldón (2009)).

Chapter 11

Future work

The work presented in this thesis evolved succesfully in a number of directions. Here we list a list of some of the open questions and extensions to the work presented in this thesis.

1 Extending the spectrum of the models available in `SPIn` and `Empar`. In particular, the models of most interest are the Algebraic Time reversible and the Stable Base Distribution models (Allman and Rhodes, 2006b). We have been able to compute the generators of the ideal for $n = 3$ taxon star tree for the `ATR` and the `SBD` model (see Ex. 5.33 in section 5.3). For a general n, however, the set of generators is unknown. The objective is to find the generating set for the spaces of all mixtures for these models. Another interesting model is the covarion model first introduced by (Tuffley and Steel, 1998; Galtier, 2001b). This model and its variants have been studied extensively (Nagaki et al., 2004; Ané et al., 2005; Galtier, 2001a; Penny et al., 2001; Misof et al., 2002; Gaucher et al., 2001; Huelsenbeck, 2002; Guindon et al., 2004) and provide a framework for modeling heterotachy.

2 One of the future goals is to provide the user with valuable information on whether the data evolved along a mixture on different tree topologies, a mixture on the same topology or from a single tree. We expect that phylogenetic invariants (although in this case they cease to be linear) can be used for this purpose. At this point, however, only a few invariants are known for these cases (see e.g. Allman et al., 2010), and further development of mathematical tools is required (see Rhodes and Sullivant (2011)).

5 In certain analyses (e.g. highly divergent sequences), working with protein alignments is preferable. A very interesting direction to pursue is the extension of the methods propsed here for DNA models to the protein coding models (Goldman and Yang, 1994). This alternative class of evolutionary models are used for modeling protein evolution and describe the amino acid replacement. Markov process has 20 states and for the reason and many approximations are being made for the analysis to be possible, i.e. the relative frequency of amino acid changes are estimated prior to inference (Dayhoff et al. (1978); Adachi and Hasegawa (1996); Jones et al. (1992); Whelan and Goldman (2001)). As in the DNA context, currently model selection approaches are defined in a continuous-time setting and use the approximated or the MLE tree (Goldman and Yang, 1994; Abascal et al., 2005).

Appendix A

Linear part of the ideal for the GMM and ATR variety

```
// GMM on b=4 states and n=4 leaves.

// f1, J1 correspond to the tree >--< with labels (12-34)
// f2, J2 correspond to the tree >--< with labels (13-24)
// f3, J3 correspond to the tree >--< with labels (14-23)
// f4, J4 correspond to the star tree.

LIB "elim.lib";



int b=4;
ring r1 = 0,(p(1..b)(1..b)(1..b)(1..b)),dp;
ring r2 = 0,(m1(1..b)(1..b),m2(1..b)(1..b),m3(1..b)(1..b),
      m4(1..b)(1..b),mi(1..b)(1..b),r(1..b),q(1..b)(1..b)(1..b)(1..b)),dp;

int i,j,k,l,u,v,s;
poly p1,p2,p3,p4;
list L1,L2,L3,L4,Q1;
s = 1;
// loop on the states at the leaves
for (i=1; i<=b; i=i+1)
{
  for(j=1; j<=b; j=j+1)
  {
    for(k=1; k<=b; k=k+1)
    {
      for(l=1; l<=b; l=l+1)
      {
        // loop for the sum in >--< (12)(34)
        p1=0;
        for(u=1; u<=b; u=u+1)
        {
```

```
        for(v=1; v<=b; v=v+1)
        {
            p1 = p1 + r(u)*m1(u)(i)*m2(u)(j)*mi(u)(v)*m3(v)(k)*m4(v)(l);

        }
      }
      L1[s] = p1;
      Q1[s] = q(i)(j)(k)(l)^6 - p1;
      // loop for the sum in >--< (13)(24)
      p2=0;
      for(u=1; u<=b; u=u+1)
      {
        for(v=1; v<=b; v=v+1)
        {
            p2 = p2 + r(u)*m1(u)(i)*m2(u)(k)*mi(u)(v)*m3(v)(j)*m4(v)(l);
        }
      }
      L2[s] = p2;


      // loop for the sum in >--< (14)(23)
      p3=0;
      for(u=1; u<=b; u=u+1)
      {
        for(v=1; v<=b; v=v+1)
        {
            p3 = p3 + r(u)*m1(u)(i)*m2(u)(l)*mi(u)(v)*m3(v)(j)*m4(v)(k);
        }
      }
      L3[s] = p3;

      // loop for the sum in the star tree
      p4=0;
      for(u=1; u<=b; u=u+1)
      {
        p4 = p4 + r(u)*m1(u)(i)*m2(u)(j)*m3(u)(k)*m4(u)(l);
      }
      L4[s] = p4;

      s = s+1;
    }
  }
}
```

```
}

map f1=r1,L1[1..b^4];
map f2=r1,L2[1..b^4];
map f3=r1,L3[1..b^4];
map f4=r1,L4[1..b^4];

// The 0 ideal
ideal I0=0;

setring r2;

// Too computationally expensive :
// ideal J1=preimage(r2,f1,I0);
// ideal J2=preimage(r2,f2,I0);
// ideal J3=preimage(r2,f2,I0);
// ideal J4=preimage(r2,f2,I0);

ideal t0=Q1[1..b^4];
degBound=6;
ideal g=std(t0); // standard Grobner basis
nselect(g,1..84);
```

We fix an order on the branches $\{0,1,2\}$ and use it to indexed the transition matrices. The first equations correspond to the stochastic condition on the transition matrices and of the root ditribution (polynomials p and d). Polynomials indexed by letter e, correspond to the condition the commutativity condition. For instance, $e121$ denotes the $(1,1)$ entry of the matrix $A^e_1A^e_2 - A^e_2A^e_1$, $e122$ is the $(1,2)$ entry, etc. The bloks are labeled by pairs $(0,1),(0,2),(1,2)$. Further, the polynomials named by f come from the condition that $D_\pi A^e$ is symmetric (6 conditions per matrix). In order to compute these polynomials we must ensure that the root distribution does not include zeros (diagonal entries of D_π are positive). Those conditions are encoded in the polynomials *root* We used variable z to homogenize the polynomials. Te goal is to express the generators of the ideal in the model parameters (the $q's$) and we choose them using the function *nselect*.

```
LIB "elim.lib";

ring r=0,(z,w(1..4),t(1..52),q(1..64)),dp;

poly p1=q(1)^5-z*(t(49)*t(1)*t(17)*t(33)+t(50)*t(5)*t(21)*t(37)+t(51)*t(9)*t(25)*t(41)+t(52)*t(13)*t(29)*t(45));
poly p2=q(2)^5-z*(t(49)*t(1)*t(17)*t(34)+t(50)*t(5)*t(21)*t(38)+t(51)*t(9)*t(25)*t(42)+t(52)*t(13)*t(29)*t(46));
poly p3=q(3)^5-z*(t(49)*t(1)*t(17)*t(35)+t(50)*t(5)*t(21)*t(39)+t(51)*t(9)*t(25)*t(43)+t(52)*t(13)*t(29)*t(47));
poly p4=q(4)^5-z*(t(49)*t(1)*t(17)*t(36)+t(50)*t(5)*t(21)*t(40)+t(51)*t(9)*t(25)*t(44)+t(52)*t(13)*t(29)*t(48));
poly p5=q(5)^5-z*(t(49)*t(1)*t(18)*t(33)+t(50)*t(5)*t(22)*t(37)+t(51)*t(9)*t(26)*t(41)+t(52)*t(13)*t(30)*t(45));
poly p6=q(6)^5-z*(t(49)*t(1)*t(18)*t(34)+t(50)*t(5)*t(22)*t(38)+t(51)*t(9)*t(26)*t(42)+t(52)*t(13)*t(30)*t(46));
poly p7=q(7)^5-z*(t(49)*t(1)*t(18)*t(35)+t(50)*t(5)*t(22)*t(39)+t(51)*t(9)*t(26)*t(43)+t(52)*t(13)*t(30)*t(47));
poly p8=q(8)^5-z*(t(49)*t(1)*t(18)*t(36)+t(50)*t(5)*t(22)*t(40)+t(51)*t(9)*t(26)*t(44)+t(52)*t(13)*t(30)*t(48));
poly p9=q(9)^5-z*(t(49)*t(1)*t(19)*t(33)+t(50)*t(5)*t(23)*t(37)+t(51)*t(9)*t(27)*t(41)+t(52)*t(13)*t(31)*t(45));
poly p10=q(10)^5-z*(t(49)*t(1)*t(19)*t(34)+t(50)*t(5)*t(23)*t(38)+t(51)*t(9)*t(27)*t(42)+t(52)*t(13)*t(31)*t(46));
poly p11=q(11)^5-z*(t(49)*t(1)*t(19)*t(35)+t(50)*t(5)*t(23)*t(39)+t(51)*t(9)*t(27)*t(43)+t(52)*t(13)*t(31)*t(47));
poly p12=q(12)^5-z*(t(49)*t(1)*t(19)*t(36)+t(50)*t(5)*t(23)*t(40)+t(51)*t(9)*t(27)*t(44)+t(52)*t(13)*t(31)*t(48));
poly p13=q(13)^5-z*(t(49)*t(1)*t(20)*t(33)+t(50)*t(5)*t(24)*t(37)+t(51)*t(9)*t(28)*t(41)+t(52)*t(13)*t(32)*t(45));
poly p14=q(14)^5-z*(t(49)*t(1)*t(20)*t(34)+t(50)*t(5)*t(24)*t(38)+t(51)*t(9)*t(28)*t(42)+t(52)*t(13)*t(32)*t(46));
poly p15=q(15)^5-z*(t(49)*t(1)*t(20)*t(35)+t(50)*t(5)*t(24)*t(39)+t(51)*t(9)*t(28)*t(43)+t(52)*t(13)*t(32)*t(47));
poly p16=q(16)^5-z*(t(49)*t(1)*t(20)*t(36)+t(50)*t(5)*t(24)*t(40)+t(51)*t(9)*t(28)*t(44)+t(52)*t(13)*t(32)*t(48));
poly p17=q(17)^5-z*(t(49)*t(2)*t(17)*t(33)+t(50)*t(6)*t(21)*t(37)+t(51)*t(10)*t(25)*t(41)+t(52)*t(14)*t(29)*t(45));
poly p18=q(18)^5-z*(t(49)*t(2)*t(17)*t(34)+t(50)*t(6)*t(21)*t(38)+t(51)*t(10)*t(25)*t(42)+t(52)*t(14)*t(29)*t(46));
poly p19=q(19)^5-z*(t(49)*t(2)*t(17)*t(35)+t(50)*t(6)*t(21)*t(39)+t(51)*t(10)*t(25)*t(43)+t(52)*t(14)*t(29)*t(47));
poly p20=q(20)^5-z*(t(49)*t(2)*t(17)*t(36)+t(50)*t(6)*t(21)*t(40)+t(51)*t(10)*t(25)*t(44)+t(52)*t(14)*t(29)*t(48));
poly p21=q(21)^5-z*(t(49)*t(2)*t(18)*t(33)+t(50)*t(6)*t(22)*t(37)+t(51)*t(10)*t(26)*t(41)+t(52)*t(14)*t(30)*t(45));
poly p22=q(22)^5-z*(t(49)*t(2)*t(18)*t(34)+t(50)*t(6)*t(22)*t(38)+t(51)*t(10)*t(26)*t(42)+t(52)*t(14)*t(30)*t(46));
poly p23=q(23)^5-z*(t(49)*t(2)*t(18)*t(35)+t(50)*t(6)*t(22)*t(39)+t(51)*t(10)*t(26)*t(43)+t(52)*t(14)*t(30)*t(47));
poly p24=q(24)^5-z*(t(49)*t(2)*t(18)*t(36)+t(50)*t(6)*t(22)*t(40)+t(51)*t(10)*t(26)*t(44)+t(52)*t(14)*t(30)*t(48));
poly p25=q(25)^5-z*(t(49)*t(2)*t(19)*t(33)+t(50)*t(6)*t(23)*t(37)+t(51)*t(10)*t(27)*t(41)+t(52)*t(14)*t(31)*t(45));
poly p26=q(26)^5-z*(t(49)*t(2)*t(19)*t(34)+t(50)*t(6)*t(23)*t(38)+t(51)*t(10)*t(27)*t(42)+t(52)*t(14)*t(31)*t(46));
poly p27=q(27)^5-z*(t(49)*t(2)*t(19)*t(35)+t(50)*t(6)*t(23)*t(39)+t(51)*t(10)*t(27)*t(43)+t(52)*t(14)*t(31)*t(47));
poly p28=q(28)^5-z*(t(49)*t(2)*t(19)*t(36)+t(50)*t(6)*t(23)*t(40)+t(51)*t(10)*t(27)*t(44)+t(52)*t(14)*t(31)*t(48));
poly p29=q(29)^5-z*(t(49)*t(2)*t(20)*t(33)+t(50)*t(6)*t(24)*t(37)+t(51)*t(10)*t(28)*t(41)+t(52)*t(14)*t(32)*t(45));
poly p30=q(30)^5-z*(t(49)*t(2)*t(20)*t(34)+t(50)*t(6)*t(24)*t(38)+t(51)*t(10)*t(28)*t(42)+t(52)*t(14)*t(32)*t(46));
poly p31=q(31)^5-z*(t(49)*t(2)*t(20)*t(35)+t(50)*t(6)*t(24)*t(39)+t(51)*t(10)*t(28)*t(43)+t(52)*t(14)*t(32)*t(47));
poly p32=q(32)^5-z*(t(49)*t(2)*t(20)*t(36)+t(50)*t(6)*t(24)*t(40)+t(51)*t(10)*t(28)*t(44)+t(52)*t(14)*t(32)*t(48));
poly p33=q(33)^5-z*(t(49)*t(3)*t(17)*t(33)+t(50)*t(7)*t(21)*t(37)+t(51)*t(11)*t(25)*t(41)+t(52)*t(15)*t(29)*t(45));
poly p34=q(34)^5-z*(t(49)*t(3)*t(17)*t(34)+t(50)*t(7)*t(21)*t(38)+t(51)*t(11)*t(25)*t(42)+t(52)*t(15)*t(29)*t(46));
poly p35=q(35)^5-z*(t(49)*t(3)*t(17)*t(35)+t(50)*t(7)*t(21)*t(39)+t(51)*t(11)*t(25)*t(43)+t(52)*t(15)*t(29)*t(47));
poly p36=q(36)^5-z*(t(49)*t(3)*t(17)*t(36)+t(50)*t(7)*t(21)*t(40)+t(51)*t(11)*t(25)*t(44)+t(52)*t(15)*t(29)*t(48));
poly p37=q(37)^5-z*(t(49)*t(3)*t(18)*t(33)+t(50)*t(7)*t(22)*t(37)+t(51)*t(11)*t(26)*t(41)+t(52)*t(15)*t(30)*t(45));
poly p38=q(38)^5-z*(t(49)*t(3)*t(18)*t(34)+t(50)*t(7)*t(22)*t(38)+t(51)*t(11)*t(26)*t(42)+t(52)*t(15)*t(30)*t(46));
poly p39=q(39)^5-z*(t(49)*t(3)*t(18)*t(35)+t(50)*t(7)*t(22)*t(39)+t(51)*t(11)*t(26)*t(43)+t(52)*t(15)*t(30)*t(47));
poly p40=q(40)^5-z*(t(49)*t(3)*t(18)*t(36)+t(50)*t(7)*t(22)*t(40)+t(51)*t(11)*t(26)*t(44)+t(52)*t(15)*t(30)*t(48));
poly p41=q(41)^5-z*(t(49)*t(3)*t(19)*t(33)+t(50)*t(7)*t(23)*t(37)+t(51)*t(11)*t(27)*t(41)+t(52)*t(15)*t(31)*t(45));
poly p42=q(42)^5-z*(t(49)*t(3)*t(19)*t(34)+t(50)*t(7)*t(23)*t(38)+t(51)*t(11)*t(27)*t(42)+t(52)*t(15)*t(31)*t(46));
poly p43=q(43)^5-z*(t(49)*t(3)*t(19)*t(35)+t(50)*t(7)*t(23)*t(39)+t(51)*t(11)*t(27)*t(43)+t(52)*t(15)*t(31)*t(47));
poly p44=q(44)^5-z*(t(49)*t(3)*t(19)*t(36)+t(50)*t(7)*t(23)*t(40)+t(51)*t(11)*t(27)*t(44)+t(52)*t(15)*t(31)*t(48));
poly p45=q(45)^5-z*(t(49)*t(3)*t(20)*t(33)+t(50)*t(7)*t(24)*t(37)+t(51)*t(11)*t(28)*t(41)+t(52)*t(15)*t(32)*t(45));
poly p46=q(46)^5-z*(t(49)*t(3)*t(20)*t(34)+t(50)*t(7)*t(24)*t(38)+t(51)*t(11)*t(28)*t(42)+t(52)*t(15)*t(32)*t(46));
poly p47=q(47)^5-z*(t(49)*t(3)*t(20)*t(35)+t(50)*t(7)*t(24)*t(39)+t(51)*t(11)*t(28)*t(43)+t(52)*t(15)*t(32)*t(47));
poly p48=q(48)^5-z*(t(49)*t(3)*t(20)*t(36)+t(50)*t(7)*t(24)*t(40)+t(51)*t(11)*t(28)*t(44)+t(52)*t(15)*t(32)*t(48));
poly p49=q(49)^5-z*(t(49)*t(4)*t(17)*t(33)+t(50)*t(8)*t(21)*t(37)+t(51)*t(12)*t(25)*t(41)+t(52)*t(16)*t(29)*t(45));
poly p50=q(50)^5-z*(t(49)*t(4)*t(17)*t(34)+t(50)*t(8)*t(21)*t(38)+t(51)*t(12)*t(25)*t(42)+t(52)*t(16)*t(29)*t(46));
poly p51=q(51)^5-z*(t(49)*t(4)*t(17)*t(35)+t(50)*t(8)*t(21)*t(39)+t(51)*t(12)*t(25)*t(43)+t(52)*t(16)*t(29)*t(47));
poly p52=q(52)^5-z*(t(49)*t(4)*t(17)*t(36)+t(50)*t(8)*t(21)*t(40)+t(51)*t(12)*t(25)*t(44)+t(52)*t(16)*t(29)*t(48));
poly p53=q(53)^5-z*(t(49)*t(4)*t(18)*t(33)+t(50)*t(8)*t(22)*t(37)+t(51)*t(12)*t(26)*t(41)+t(52)*t(16)*t(30)*t(45));
poly p54=q(54)^5-z*(t(49)*t(4)*t(18)*t(34)+t(50)*t(8)*t(22)*t(38)+t(51)*t(12)*t(26)*t(42)+t(52)*t(16)*t(30)*t(46));
poly p55=q(55)^5-z*(t(49)*t(4)*t(18)*t(35)+t(50)*t(8)*t(22)*t(39)+t(51)*t(12)*t(26)*t(43)+t(52)*t(16)*t(30)*t(47));
poly p56=q(56)^5-z*(t(49)*t(4)*t(18)*t(36)+t(50)*t(8)*t(22)*t(40)+t(51)*t(12)*t(26)*t(44)+t(52)*t(16)*t(30)*t(48));
poly p57=q(57)^5-z*(t(49)*t(4)*t(19)*t(33)+t(50)*t(8)*t(23)*t(37)+t(51)*t(12)*t(27)*t(41)+t(52)*t(16)*t(31)*t(45));
poly p58=q(58)^5-z*(t(49)*t(4)*t(19)*t(34)+t(50)*t(8)*t(23)*t(38)+t(51)*t(12)*t(27)*t(42)+t(52)*t(16)*t(31)*t(46));
poly p59=q(59)^5-z*(t(49)*t(4)*t(19)*t(35)+t(50)*t(8)*t(23)*t(39)+t(51)*t(12)*t(27)*t(43)+t(52)*t(16)*t(31)*t(47));
poly p60=q(60)^5-z*(t(49)*t(4)*t(19)*t(36)+t(50)*t(8)*t(23)*t(40)+t(51)*t(12)*t(27)*t(44)+t(52)*t(16)*t(31)*t(48));
poly p61=q(61)^5-z*(t(49)*t(4)*t(20)*t(33)+t(50)*t(8)*t(24)*t(37)+t(51)*t(12)*t(28)*t(41)+t(52)*t(16)*t(32)*t(45));
poly p62=q(62)^5-z*(t(49)*t(4)*t(20)*t(34)+t(50)*t(8)*t(24)*t(38)+t(51)*t(12)*t(28)*t(42)+t(52)*t(16)*t(32)*t(46));
poly p63=q(63)^5-z*(t(49)*t(4)*t(20)*t(35)+t(50)*t(8)*t(24)*t(39)+t(51)*t(12)*t(28)*t(43)+t(52)*t(16)*t(32)*t(47));
poly p64=q(64)^5-z*(t(49)*t(4)*t(20)*t(36)+t(50)*t(8)*t(24)*t(40)+t(51)*t(12)*t(28)*t(44)+t(52)*t(16)*t(32)*t(48));


poly d1=t(1)+t(2)+t(3)+t(4)-z;
poly d2=t(5)+t(6)+t(7)+t(8)-z;
poly d3=t(9)+t(10)+t(11)+t(12)-z;
poly d4=t(13)+t(14)+t(15)+t(16)-z;
poly d5=t(17)+t(18)+t(19)+t(20)-z;
poly d6=t(21)+t(22)+t(23)+t(24)-z;
poly d7=t(25)+t(26)+t(27)+t(28)-z;
poly d8=t(29)+t(30)+t(31)+t(32)-z;
poly d9=t(33)+t(34)+t(35)+t(36)-z;
poly d10=t(37)+t(38)+t(39)+t(40)-z;
poly d11=t(41)+t(42)+t(43)+t(44)-z;
poly d12=t(45)+t(46)+t(47)+t(48)-z;


poly e011=t(1)*t(17)+t(2)*t(21)+t(3)*t(25)+t(4)*t(29)-t(17)*t(1)-t(18)*t(5)-t(19)*t(9)-t(20)*t(13);
poly e012=t(1)*t(18)+t(2)*t(22)+t(3)*t(26)+t(4)*t(30)-t(17)*t(2)-t(18)*t(6)-t(19)*t(10)-t(20)*t(14);
poly e013=t(1)*t(19)+t(2)*t(23)+t(3)*t(27)+t(4)*t(31)-t(17)*t(3)-t(18)*t(7)-t(19)*t(11)-t(20)*t(15);
poly e014=t(1)*t(20)+t(2)*t(24)+t(3)*t(28)+t(4)*t(32)-t(17)*t(4)-t(18)*t(8)-t(19)*t(12)-t(20)*t(16);
poly e015=t(5)*t(17)+t(6)*t(21)+t(7)*t(25)+t(8)*t(29)-t(21)*t(1)-t(22)*t(5)-t(23)*t(9)-t(24)*t(13);
poly e016=t(5)*t(18)+t(6)*t(22)+t(7)*t(26)+t(8)*t(30)-t(21)*t(2)-t(22)*t(6)-t(23)*t(10)-t(24)*t(14);
poly e017=t(5)*t(19)+t(6)*t(23)+t(7)*t(27)+t(8)*t(31)-t(21)*t(3)-t(22)*t(7)-t(23)*t(11)-t(24)*t(15);
poly e018=t(5)*t(20)+t(6)*t(24)+t(7)*t(28)+t(8)*t(32)-t(21)*t(4)-t(22)*t(8)-t(23)*t(12)-t(24)*t(16);
poly e019=t(9)*t(17)+t(10)*t(21)+t(11)*t(25)+t(12)*t(29)-t(25)*t(1)-t(26)*t(5)-t(27)*t(9)-t(28)*t(13);
poly e0110=t(9)*t(18)+t(10)*t(22)+t(11)*t(26)+t(12)*t(30)-t(25)*t(2)-t(26)*t(6)-t(27)*t(10)-t(28)*t(14);
poly e0111=t(9)*t(19)+t(10)*t(23)+t(11)*t(27)+t(12)*t(31)-t(25)*t(3)-t(26)*t(7)-t(27)*t(11)-t(28)*t(15);
poly e0112=t(9)*t(20)+t(10)*t(24)+t(11)*t(28)+t(12)*t(32)-t(25)*t(4)-t(26)*t(8)-t(27)*t(12)-t(28)*t(16);
poly e0113=t(13)*t(17)+t(14)*t(21)+t(15)*t(25)+t(16)*t(29)-t(29)*t(1)-t(30)*t(5)-t(31)*t(9)-t(32)*t(13);
poly e0114=t(13)*t(18)+t(14)*t(22)+t(15)*t(26)+t(16)*t(30)-t(29)*t(2)-t(30)*t(6)-t(31)*t(10)-t(32)*t(14);
poly e0115=t(13)*t(19)+t(14)*t(23)+t(15)*t(27)+t(16)*t(31)-t(29)*t(3)-t(30)*t(7)-t(31)*t(11)-t(32)*t(15);
poly e0116=t(13)*t(20)+t(14)*t(24)+t(15)*t(28)+t(16)*t(32)-t(29)*t(4)-t(30)*t(8)-t(31)*t(12)-t(32)*t(16);
poly e021=t(1)*t(33)+t(2)*t(37)+t(3)*t(41)+t(4)*t(45)-t(33)*t(1)-t(34)*t(5)-t(35)*t(9)-t(36)*t(13);
poly e022=t(1)*t(34)+t(2)*t(38)+t(3)*t(42)+t(4)*t(46)-t(33)*t(2)-t(34)*t(6)-t(35)*t(10)-t(36)*t(14);
poly e023=t(1)*t(35)+t(2)*t(39)+t(3)*t(43)+t(4)*t(47)-t(33)*t(3)-t(34)*t(7)-t(35)*t(11)-t(36)*t(15);
poly e024=t(1)*t(36)+t(2)*t(40)+t(3)*t(44)+t(4)*t(48)-t(33)*t(4)-t(34)*t(8)-t(35)*t(12)-t(36)*t(16);
poly e025=t(5)*t(33)+t(6)*t(37)+t(7)*t(41)+t(8)*t(45)-t(37)*t(1)-t(38)*t(5)-t(39)*t(9)-t(40)*t(13);
poly e026=t(5)*t(34)+t(6)*t(38)+t(7)*t(42)+t(8)*t(46)-t(37)*t(2)-t(38)*t(6)-t(39)*t(10)-t(40)*t(14);
poly e027=t(5)*t(35)+t(6)*t(39)+t(7)*t(43)+t(8)*t(47)-t(37)*t(3)-t(38)*t(7)-t(39)*t(11)-t(40)*t(15);
poly e028=t(5)*t(36)+t(6)*t(40)+t(7)*t(44)+t(8)*t(48)-t(37)*t(4)-t(38)*t(8)-t(39)*t(12)-t(40)*t(16);
poly e029=t(9)*t(33)+t(10)*t(37)+t(11)*t(41)+t(12)*t(45)-t(41)*t(1)-t(42)*t(5)-t(43)*t(9)-t(44)*t(13);
poly e0210=t(9)*t(34)+t(10)*t(38)+t(11)*t(42)+t(12)*t(46)-t(41)*t(2)-t(42)*t(6)-t(43)*t(10)-t(44)*t(14);
poly e0211=t(9)*t(35)+t(10)*t(39)+t(11)*t(43)+t(12)*t(47)-t(41)*t(3)-t(42)*t(7)-t(43)*t(11)-t(44)*t(15);
poly e0212=t(9)*t(36)+t(10)*t(40)+t(11)*t(44)+t(12)*t(48)-t(41)*t(4)-t(42)*t(8)-t(43)*t(12)-t(44)*t(16);
poly e0213=t(13)*t(33)+t(14)*t(37)+t(15)*t(41)+t(16)*t(45)-t(45)*t(1)-t(46)*t(5)-t(47)*t(9)-t(48)*t(13);
poly e0214=t(13)*t(34)+t(14)*t(38)+t(15)*t(42)+t(16)*t(46)-t(45)*t(2)-t(46)*t(6)-t(47)*t(10)-t(48)*t(14);
poly e0215=t(13)*t(35)+t(14)*t(39)+t(15)*t(43)+t(16)*t(47)-t(45)*t(3)-t(46)*t(7)-t(47)*t(11)-t(48)*t(15);
poly e0216=t(13)*t(36)+t(14)*t(40)+t(15)*t(44)+t(16)*t(48)-t(45)*t(4)-t(46)*t(8)-t(47)*t(12)-t(48)*t(16);

poly e121=t(17)*t(33)+t(18)*t(37)+t(19)*t(41)+t(20)*t(45)-t(33)*t(17)-t(34)*t(21)-t(35)*t(25)-t(36)*t(29);
poly e122=t(17)*t(34)+t(18)*t(38)+t(19)*t(42)+t(20)*t(46)-t(33)*t(18)-t(34)*t(22)-t(35)*t(26)-t(36)*t(30);
poly e123=t(17)*t(35)+t(18)*t(39)+t(19)*t(43)+t(20)*t(47)-t(33)*t(19)-t(34)*t(23)-t(35)*t(27)-t(36)*t(31);
poly e124=t(17)*t(36)+t(18)*t(40)+t(19)*t(44)+t(20)*t(48)-t(33)*t(20)-t(34)*t(24)-t(35)*t(28)-t(36)*t(32);
poly e125=t(21)*t(33)+t(22)*t(37)+t(23)*t(41)+t(24)*t(45)-t(37)*t(17)-t(38)*t(21)-t(39)*t(25)-t(40)*t(29);
poly e126=t(21)*t(34)+t(22)*t(38)+t(23)*t(42)+t(24)*t(46)-t(37)*t(18)-t(38)*t(22)-t(39)*t(26)-t(40)*t(30);
poly e127=t(21)*t(35)+t(22)*t(39)+t(23)*t(43)+t(24)*t(47)-t(37)*t(19)-t(38)*t(23)-t(39)*t(27)-t(40)*t(31);
poly e128=t(21)*t(36)+t(22)*t(40)+t(23)*t(44)+t(24)*t(48)-t(37)*t(20)-t(38)*t(24)-t(39)*t(28)-t(40)*t(32);
poly e129=t(25)*t(33)+t(26)*t(37)+t(27)*t(41)+t(28)*t(45)-t(41)*t(17)-t(42)*t(21)-t(43)*t(25)-t(44)*t(29);
poly e1210=t(25)*t(34)+t(26)*t(38)+t(27)*t(42)+t(28)*t(46)-t(41)*t(18)-t(42)*t(22)-t(43)*t(26)-t(44)*t(30);
poly e1211=t(25)*t(35)+t(26)*t(39)+t(27)*t(43)+t(28)*t(47)-t(41)*t(19)-t(42)*t(23)-t(43)*t(27)-t(44)*t(31);
poly e1212=t(25)*t(36)+t(26)*t(40)+t(27)*t(44)+t(28)*t(48)-t(41)*t(20)-t(42)*t(24)-t(43)*t(28)-t(44)*t(32);
poly e1213=t(29)*t(33)+t(30)*t(37)+t(31)*t(41)+t(32)*t(45)-t(45)*t(17)-t(46)*t(21)-t(47)*t(25)-t(48)*t(29);
poly e1214=t(29)*t(34)+t(30)*t(38)+t(31)*t(42)+t(32)*t(46)-t(45)*t(18)-t(46)*t(22)-t(47)*t(26)-t(48)*t(30);
poly e1215=t(29)*t(35)+t(30)*t(39)+t(31)*t(43)+t(32)*t(47)-t(45)*t(19)-t(46)*t(23)-t(47)*t(27)-t(48)*t(31);
poly e1216=t(29)*t(36)+t(30)*t(40)+t(31)*t(44)+t(32)*t(48)-t(45)*t(20)-t(46)*t(24)-t(47)*t(28)-t(48)*t(32);


poly f01=t(49)*t(2) - t(50)*t(5);
poly f02=t(49)*t(3) - t(51)*t(9);
poly f03=t(49)*t(4) - t(52)*t(13);
poly f04=t(50)*t(7) - t(51)*t(10);
poly f05=t(50)*t(8) - t(52)*t(14);
poly f06=t(51)*t(12) - t(52)*t(15);

poly f11=t(49)*t(18) - t(50)*t(21);
poly f12=t(49)*t(19) - t(51)*t(25);
poly f13=t(49)*t(20) - t(52)*t(29);
poly f14=t(50)*t(23) - t(51)*t(26);
poly f15=t(50)*t(24) - t(52)*t(30);
poly f16=t(51)*t(28) - t(52)*t(31);
poly f21=t(49)*t(34) - t(50)*t(37);
poly f22=t(49)*t(35) - t(51)*t(41);
poly f23=t(49)*t(36) - t(52)*t(45);
poly f24=t(50)*t(39) - t(51)*t(42);
poly f25=t(50)*t(40) - t(52)*t(46);
poly f26=t(51)*t(44) - t(52)*t(47);

poly root1=w(1)*t(49) - z^2;
poly root2=w(2)*t(50) - z^2;
poly root3=w(3)*t(51) - z^2;
poly root4=w(4)*t(52) - z^2;

ideal t0= d1, d2, d3, d4, d5, d6, d7, d8, d9, d10, d11, d12, e011, e012, e013, e014, e015, e016, e017, e018, e019,
          e0110, e0111, e0112, e0113, e0114, e0115, e0116, e021, e022, e023, e024, e025, e026, e027, e028, e029,
          e0210, e0211, e0212, e0213, e0214, e0215, e0216, e121, e122, e123, e124, e125, e126, e127, e128, e129,
          e1210, e1211, e1212, e1213, e1214, e1215, e1216, f01, f02, f03, f04, f05, f06, f11, f12, f13, f14, f15,
          f16, f21, f22, f23, f24, f25, f26, root1,root2, root3, root4, p1, p2, p3, p4, p5, p6, p7, p8, p9, p10,
          p11, p12, p13, p14, p15, p16, p17, p18, p19, p20, p21, p22, p23, p24, p25, p26, p27, p28, p29, p30, p31,
          p32, p33, p34, p35, p36, p37,  p38,  p39, p40, p41, p42, p43, p44, p45, p46, p47, p48, p49, p50, p51,p52,
          p53, p54, p55, p56, p57, p58, p59, p60, p61, p62, p63, p64;
degBound=5;
ideal g=std(t0);
nselect(g,1,57);
```

Here we give the output of the above code :

```
                    SINGULAR                             /  Development
 A Computer Algebra System for Polynomial Computations   /   version 3-0-4
                                                       0<
    by: G.-M. Greuel, G. Pfister, H. Schoenemann        \   Nov 2007
FB Mathematik der Universitaet, D-67653 Kaiserslautern    \
// ** loaded /usr/lib/singular/elim.lib (1.22,2008/04/22)
// ** loaded /usr/lib/singular/poly.lib (1.46,2007/07/25)
// ** loaded /usr/lib/singular/ring.lib (1.32,2008/03/25)
// ** loaded /usr/lib/singular/primdec.lib (1.139,2008/03/19)
// ** loaded /usr/lib/singular/absfact.lib (1.6,2007/07/13)
// ** loaded /usr/lib/singular/triang.lib (1.11,2006/12/06)
// ** loaded /usr/lib/singular/matrix.lib (1.41,2007/12/22)
// ** loaded /usr/lib/singular/random.lib (1.17,2006/07/20)
// ** loaded /usr/lib/singular/general.lib (1.56,2008/03/18)
// ** loaded /usr/lib/singular/inout.lib (1.30,2007/11/29)
_[1]=q(45)^5+q(46)^5+q(47)^5+q(48)^5-q(57)^5-q(58)^5-q(59)^5-q(60)^5
_[2]=q(36)^5+q(40)^5+q(44)^5+q(48)^5-q(51)^5-q(55)^5-q(59)^5-q(63)^5
_[3]=q(29)^5+q(30)^5+q(31)^5+q(32)^5-q(53)^5-q(54)^5-q(55)^5-q(56)^5
_[4]=q(25)^5+q(26)^5+q(27)^5+q(28)^5-q(37)^5-q(38)^5-q(39)^5-q(40)^5
_[5]=q(20)^5+q(24)^5+q(28)^5+q(32)^5-q(50)^5-q(54)^5-q(58)^5-q(62)^5
_[6]=q(19)^5+q(23)^5+q(27)^5+q(31)^5-q(34)^5-q(38)^5-q(42)^5-q(46)^5
_[7]=q(13)^5+q(14)^5+q(15)^5+q(16)^5-q(49)^5-q(50)^5-q(51)^5-q(52)^5
_[8]=q(12)^5-q(15)^5+q(28)^5-q(31)^5+q(44)^5-q(47)^5+q(60)^5-q(63)^5
_[9]=q(9)^5+q(10)^5+q(11)^5+q(15)^5-q(28)^5+q(31)^5-q(33)^5-q(34)^5
    -q(35)^5+q(40)^5+q(47)^5+q(48)^5-q(51)^5-q(55)^5-q(59)^5-q(60)^5
_[10]=q(8)^5-q(14)^5+q(24)^5-q(30)^5+q(40)^5-q(46)^5+q(56)^5-q(62)^5
_[11]=q(7)^5-q(10)^5+q(23)^5-q(26)^5+q(39)^5-q(42)^5+q(55)^5-q(58)^5
_[12]=q(5)^5+q(6)^5+q(10)^5+q(14)^5-q(17)^5-q(18)^5+q(26)^5+q(27)^5
     +q(28)^5+q(30)^5+q(31)^5+q(32)^5-q(34)^5-q(38)^5-q(39)^5-q(40)^5
     -q(50)^5-q(54)^5-q(55)^5-q(56)^5
_[13]=q(4)^5-q(13)^5-q(24)^5-q(28)^5-q(29)^5-q(32)^5-q(40)^5-q(44)^5
     -q(45)^5-q(48)^5+q(50)^5+q(51)^5+q(52)^5+q(54)^5+q(55)^5+q(58)^5
     +q(59)^5-q(61)^5+q(62)^5+q(63)^5
_[14]=q(3)^5-q(9)^5-q(23)^5-q(25)^5-q(27)^5-q(31)^5+q(34)^5+q(35)^5
     +q(38)^5-q(41)^5+q(42)^5+q(46)^5+q(51)^5-q(57)^5
_[15]=q(2)^5-q(5)^5+q(18)^5-q(21)^5+q(34)^5-q(37)^5+q(50)^5-q(53)^5
```

Appendix B
Observed Fisher Information Matrix

Consider a discrete-time Markov model $\mathcal{M}$ with equal row composition, i.e. up to a permutation the set of free parameters in each row is the same. Let us denote by d the degrees of freedom of the model, so that the total number of parameters for any substitution matrix in $\mathcal{M}$ is $d+1$. In addition, let us assume that the root distribution is uniform.

First we derive the formula for the Fisher information matrix omitting the stochastic condition of matrix rows summing to 1.

Let $\mathcal{T}$ be a phylogenetic tree and let $\xi = (\xi^e_k)_{k=1,\ldots,d+1,e\in E(\mathcal{T})}$ be the vector of parameters of $\mathcal{M}$ (i.e the distinct entries of the transition matrices A^e for the edges e of $\mathcal{T}$.)

Let $\mathbf{y}$ denote a set of states assigned jointly to the hidden nodes (including the root) and $\mathbf{x}$ a pattern at the $\mathtt{L}(\mathcal{T})$, e.g. $\mathbf{x} = (\underbrace{\mathtt{a}\ldots,\mathtt{a}}_{|\mathtt{L}(\mathcal{T})|})$. Also, denote by X the set of $\mathbf{x}$. Given the states in the complete model, $(\mathbf{x},\mathbf{y})$, let $\alpha(e,\mathbf{x},\mathbf{y})$ denote the corresponding index of the parameter in A^e edge e. It is the index of the entry of A^e given by the states in the parent and child nodes of e dictated by $\mathbf{x}$ and $\mathbf{y}$. For instance, if the states at the two ends of e are $\mathtt{c}$ and $\mathtt{g}$, then $\alpha(e,\mathbf{x},\mathbf{y})$ is the index of the entry $(2,3)th$ entry of A^e. For notational convenience, let $n_e = |E(\mathcal{T})|$. Also, we write $(\mathrm{u}_{\mathbf{x}})_{\mathbf{x}\in X} = \{\mathrm{u}_{\mathtt{a}\ldots\mathtt{a}}, \mathrm{u}_{\mathtt{a}\ldots\mathtt{c}}, \ldots, \mathrm{u}_{\mathtt{t}\ldots\mathtt{t}}\}$ for the set of occurrence of the observed patterns in the columns of the alignment. We can write the formula for the joint probability of a pattern $\mathbf{x}$ as

$$p_{\mathbf{x}}(\xi) = \sum_{\mathbf{y}} \prod_{e=1}^{n_e} \xi^e_{\alpha(e,\mathbf{x},\mathbf{y})}.$$

We first list the formulas for the derivatives

$$\frac{\partial p_{\mathbf{x}}(\xi)}{\partial \xi^e_k} = \sum_{\mathbf{y}:\alpha(e,\mathbf{x},\mathbf{y})=k} \prod_{f\neq e} \xi^f_{\alpha(\mathbf{x},\mathbf{y})}$$

In the second derivatives, deriving twice with respect to the same edges, that is when $e = e'$, gives $\frac{\partial^2 p_{\mathbf{x}}(\xi)}{\partial \xi^e_k \partial \xi^e_{k'}} = 0$, since the sum of the monomials in the expression for the joint probability contains exactly one ξ^e_k per branch.

On the other hand if $e \neq e'$

$$\frac{\partial^2 p_{\mathbf{x}}(\xi)}{\partial \xi^e_k \partial \xi^{e'}_{k'}} = \sum_{\substack{J \text{ such that} \\ \alpha(e,I,J)=k \\ \alpha(e',I,J)=k'}} \prod_{f \neq e,e'} \xi^f_{\alpha(f,I,J)}$$

We will write u_D for the set of counts of patterns at the leaves $\mathtt{u}_D = (\mathtt{u}_{\mathbf{x}})_{\mathbf{x}\in X}$, which is a vector of length $4^{|\mathtt{L}(\mathcal{T})|}$. We note that for some of its entries the counts will be 0. Let $L = \sum_{\mathbf{x}\in X} \mathtt{u}_{\mathbf{x}}$.

The Fisher information matrix with the unrestricted parameters is

$$\begin{aligned}
\mathbf{I}_{\text{un}}(e,k;e',k') &= -\mathbf{E}\left(\frac{\partial^2 \mathcal{L}_{obs}(\xi;u_D)}{\partial \xi^e_k \partial \xi^{e'}_{k'}}\right) \\
&= \sum_{\mathbf{x}\in X} -\mathbf{E}(\mathtt{u}_{\mathbf{x}}) \frac{\partial^2}{\partial \xi^e_k \partial \xi^{e'}_{k'}} \log p_{\mathbf{x}}(\xi) \\
&= \sum_{\mathbf{x}\in X} -L p_{\mathbf{x}}(\xi) \frac{\partial}{\partial \xi^e_k}\left(\frac{1}{p_{\mathbf{x}}(\xi)} \frac{\partial p_{\mathbf{x}}(\xi)}{\partial \xi^{e'}_{k'}}\right) \\
&= \sum_{\mathbf{x}\in X} -L p_{\mathbf{x}}(\xi) \left(\frac{-1}{p_{\mathbf{x}}(\xi)^2} \frac{\partial p_{\mathbf{x}}(\xi)}{\partial \xi^e_k} \frac{\partial p_{\mathbf{x}}(\xi)}{\partial \xi^{e'}_{k'}} + \frac{1}{p_{\mathbf{x}}(\xi)} \frac{\partial^2 p_{\mathbf{x}}(\xi)}{\partial \xi^e_k \partial \xi^{e'}_{k'}}\right) \\
&= L \sum_{\mathbf{x}\in X} \left(\frac{1}{p_{\mathbf{x}}(\xi)} \frac{\partial p_{\mathbf{x}}(\xi)}{\partial \xi^e_k} \frac{\partial p_{\mathbf{x}}(\xi)}{\partial \xi^{e'}_{k'}} - \frac{\partial^2 p_{\mathbf{x}}(\xi)}{\partial \xi^e_k \partial \xi^{e'}_{k'}}\right)
\end{aligned}$$

We used the fact that the expected value of the sample mean is the population mean: $\mathbf{E}(\mathtt{u}_{\mathbf{x}}) = L p_{\mathbf{x}}(\xi)$. In the above derivation we used the chain rule.

As a last step we add the stochastic condition and compute the Fisher information matrix for the free parameters. We will denote this $dn_e \times dn_e$ matrix by $\mathbf{I}$. Stochastic condition is the same for each row:

$$\xi^e_1 = 1 - C_2 \xi^e_2 - \cdots - C_{d+1} \xi^e_{d+1},$$

where C_i the number of times the parameter ξ^e_i appears in a row. Note that $\xi = (\xi^e_k)_{k=2,\ldots,d+1}$ are now the free parameters. For example, for $\mathtt{K81}^* d = 3$ and $C_2 = C_3 = C_4 = 1$ and for $\mathtt{JC69}^*$ $d = 1$, so $C_2 = 3$. The particular structure of the models we consider (rows contain the same set of free parameters), and modeling the evolutionary process by the same model at distinct branches, we have that $C^e_k = C^{e'}_k$.

Now, for $k, k' = 2\ldots, d+1$, we have

$$\begin{aligned}
\mathbf{I}(\xi^e_k, \xi^{e'}_{k'}) = -\mathbf{E}\left(\frac{\partial^2 \mathcal{L}_{obs}(\xi;u_D)}{\partial \xi^e_k \partial \xi^{e'}_{k'}}\right) &= -\mathbf{E}\left(\frac{\partial}{\partial \xi^e_k}\left(\frac{\partial \mathcal{L}_{obs}(\xi;u_D)}{\partial \xi^{e'}_1} \frac{\partial \xi^{e'}_1}{\partial \xi^{e'}_{k'}} + \frac{\partial \mathcal{L}_{obs}(\xi;u_D)}{\partial \xi^{e'}_{k'}}\right)\right) \\
&= -\mathbf{E}\left(\left(\frac{\partial^2 \mathcal{L}_{obs}(\xi;u_D)}{\partial \xi^e_1 \partial \xi^{e'}_1} \frac{\partial \xi^e_1}{\partial \xi^e_k} + \frac{\partial^2 \mathcal{L}_{obs}(\xi;u_D)}{\partial \xi^e_k \partial \xi^{e'}_1}\right) \frac{\partial \xi^{e'}_1}{\partial \xi^{e'}_{k'}}\right) \\
&\quad - \mathbf{E}\left(\frac{\partial^2 \mathcal{L}_{obs}(\xi;u_D)}{\partial \xi^e_1 \partial \xi^{e'}_{k'}} \frac{\partial \xi^e_1}{\partial \xi^e_k} + \frac{\partial^2 \mathcal{L}_{obs}(\xi;u_D)}{\partial \xi^e_k \partial \xi^{e'}_{k'}}\right) = \mathbf{I}_{un}(e,1;e',1) C_k C_{k'} \\
&\quad - \mathbf{I}_{un}(e,k;e',1) C_{k'} - \mathbf{I}_{un}(e,1;e',k') C_k + \mathbf{I}_{un}(e,k;e',k').
\end{aligned}$$

The formulae for the SSM and the GMM model can be obtained analogously by adding extra stochastic conditions for the remaining rows of the matrix A^e (1 for the SSM and 3 for the GMM) and the conditions of the root distribution (stochastic condition for both models and additional base-pairing property for the SSM).

Appendix C

`Empar` performance assessement

Table C.1: The relative frequency of the χ^2 tests based on the asymptotic normality of the maximum likelihood estimator with *p-value*$\in (0.05, 0.95)$ calculated from 1.000 simulations under the `JC69`* model. Each data set was a multiple sequence alignment generated on the $\mathcal{T}_{\text{balanced}}$ tree with the depth branch length set to the values indicated by the first columns. We present results for a variety of data lengths L and with the distinction as to the positioning of the branch in the tree.

	depth 1				depth 2			
l \ L	300nt	500nt	1.000nt	10.000nt	300nt	500nt	1.000nt	10.000nt
0.01	0.968	0.970	0.973	0.947	0.979	0.978	0.971	0.955
0.05	0.955	0.949	0.952	0.941	0.933	0.959	0.950	0.945
0.10	0.953	0.943	0.947	0.951	0.967	0.953	0.938	0.965
0.15	0.951	0.950	0.957	0.951	0.949	0.945	0.947	0.944
0.20	0.946	0.969	0.947	0.947	0.938	0.936	0.949	0.957
0.25	0.949	0.948	0.935	0.951	0.941	0.945	0.956	0.942
0.30	0.944	0.959	0.946	0.960	0.948	0.945	0.942	0.951
0.35	0.956	0.950	0.946	0.938	0.947	0.954	0.958	0.948
0.40	0.940	0.949	0.948	0.947	0.949	0.934	0.945	0.946
0.45	0.954	0.942	0.944	0.956	0.925	0.943	0.947	0.959
0.50	0.948	0.949	0.950	0.957	0.945	0.945	0.957	0.949
0.55	0.939	0.942	0.953	0.948	0.948	0.939	0.946	0.953
0.60	0.919	0.940	0.938	0.946	0.926	0.944	0.933	0.940
0.65	0.887	0.919	0.928	0.949	0.933	0.918	0.951	0.955
0.70	0.890	0.890	0.928	0.948	0.935	0.926	0.931	0.940
0.75	0.864	0.884	0.920	0.953	0.950	0.939	0.930	0.957
0.80	0.862	0.890	0.898	0.958	0.952	0.937	0.927	0.943
0.85	0.845	0.844	0.890	0.949	0.937	0.958	0.921	0.949
0.90	0.821	0.818	0.874	0.939	0.928	0.943	0.948	0.944
0.95	0.767	0.806	0.848	0.943	0.913	0.955	0.952	0.935
1.00	0.788	0.784	0.820	0.942	0.902	0.930	0.952	0.939
1.05	0.757	0.784	0.800	0.907	0.877	0.933	0.952	0.924
1.10	0.778	0.785	0.805	0.894	0.970	0.894	0.943	0.900
1.15	0.968	0.771	0.776	0.862	0.967	0.861	0.929	0.895
1.20	0.960	0.794	0.760	0.865	0.959	0.875	0.923	0.915
1.25	0.956	0.962	0.754	0.839	0.956	0.967	0.903	0.961
1.30	0.950	0.960	0.777	0.836	0.942	0.960	0.870	0.971
1.35	0.933	0.956	0.800	0.796	0.942	0.954	0.836	0.967
1.40	0.929	0.948	0.963	0.783	0.944	0.959	0.966	0.965

Table C.2: The relative frequency of the χ^2 tests based on the asymptotic normality of the maximum likelihood estimator with *p-value*$\in (0.05, 0.95)$ calculated from 1.000 simulations under the `JC69`* model. Each data set was a multiple sequence alignment generated on the $\mathcal{T}_{2:1}$ tree with the depth branch length set to the values indicated by the first columns. We present results for a variety of data lengths L and with the distinction as to the positioning of the branch in the tree.

	depth 1				depth 2			
l \ L	300nt	500nt	1.000nt	10.000nt	300nt	500nt	1.000nt	10.000nt
0.01	0.990	0.965	0.979	0.940	0.962	0.968	0.957	0.942
0.05	0.955	0.938	0.959	0.946	0.955	0.944	0.939	0.961
0.10	0.953	0.958	0.963	0.950	0.948	0.950	0.950	0.952
0.15	0.961	0.959	0.939	0.959	0.962	0.945	0.948	0.946
0.20	0.950	0.944	0.953	0.940	0.953	0.945	0.943	0.947
0.25	0.953	0.950	0.952	0.941	0.954	0.966	0.955	0.946
0.30	0.957	0.944	0.955	0.950	0.942	0.939	0.965	0.951
0.35	0.940	0.952	0.949	0.943	0.945	0.953	0.965	0.953
0.40	0.922	0.959	0.951	0.947	0.940	0.950	0.954	0.948
0.45	0.946	0.958	0.949	0.954	0.947	0.949	0.957	0.942
0.50	0.942	0.958	0.946	0.943	0.957	0.941	0.942	0.953
0.55	0.960	0.937	0.954	0.942	0.966	0.944	0.950	0.951
0.60	0.939	0.950	0.947	0.959	0.955	0.958	0.958	0.950
0.65	0.942	0.946	0.949	0.937	0.941	0.956	0.945	0.939
0.70	0.953	0.934	0.944	0.941	0.943	0.941	0.953	0.953
0.75	0.937	0.937	0.946	0.956	0.951	0.958	0.944	0.945
0.80	0.931	0.942	0.929	0.943	0.936	0.951	0.948	0.951
0.85	0.914	0.920	0.939	0.943	0.945	0.948	0.929	0.947
0.90	0.904	0.910	0.947	0.934	0.943	0.937	0.936	0.945
0.95	0.899	0.921	0.929	0.957	0.949	0.933	0.944	0.953
1.00	0.918	0.911	0.941	0.956	0.949	0.925	0.950	0.949
1.05	0.898	0.901	0.924	0.953	0.954	0.943	0.942	0.950
1.10	0.880	0.895	0.918	0.941	0.955	0.957	0.942	0.963
1.15	0.875	0.901	0.908	0.933	0.958	0.957	0.924	0.945
1.20	0.859	0.871	0.895	0.954	0.951	0.959	0.944	0.957
1.25	0.853	0.859	0.905	0.964	0.955	0.950	0.938	0.940
1.30	0.826	0.839	0.883	0.935	0.952	0.961	0.944	0.939
1.35	0.810	0.829	0.878	0.942	0.947	0.957	0.964	0.962
1.40	0.974	0.836	0.875	0.930	0.952	0.962	0.951	0.932

Table C.3: The relative frequency of the χ^2 tests based on the asymptotic normality of the maximum likelihood estimator with *p-value*$\in (0.05, 0.95)$ calculated from 1.000 simulations under the `K81`* model. Each data set was a multiple sequence alignment generated on the $\mathcal{T}^4_{\text{balanced}}$ tree with the depth branch length set to the values indicated by the first columns. We present results for a variety of data lengths L and with the distinction as to the positioning of the branch in the tree.

	depth 1				depth 2			
l \ L	300nt	500nt	1.000nt	10.000nt	300nt	500nt	1.000nt	10.000nt
0.01	0.904	0.930	0.931	0.945	0.900	0.919	0.936	0.954
0.05	0.942	0.949	0.941	0.947	0.932	0.948	0.941	0.949
0.10	0.937	0.948	0.949	0.949	0.965	0.933	0.941	0.945
0.15	0.936	0.942	0.947	0.953	0.945	0.953	0.941	0.951
0.20	0.938	0.938	0.944	0.955	0.947	0.938	0.944	0.953
0.25	0.946	0.951	0.956	0.949	0.950	0.950	0.954	0.953
0.30	0.949	0.959	0.954	0.965	0.947	0.942	0.941	0.958
0.35	0.943	0.935	0.935	0.949	0.944	0.964	0.952	0.950
0.40	0.945	0.938	0.950	0.950	0.939	0.947	0.951	0.956
0.45	0.922	0.938	0.953	0.937	0.942	0.954	0.948	0.939
0.50	0.919	0.937	0.941	0.944	0.932	0.925	0.954	0.956
0.55	0.930	0.931	0.937	0.948	0.926	0.927	0.951	0.944
0.60	0.902	0.911	0.937	0.960	0.923	0.936	0.953	0.947
0.65	0.913	0.921	0.928	0.949	0.916	0.937	0.939	0.959
0.70	0.880	0.895	0.926	0.943	0.908	0.909	0.935	0.943
0.75	0.883	0.907	0.903	0.935	0.924	0.906	0.927	0.941
0.80	0.877	0.860	0.884	0.941	0.899	0.903	0.921	0.941
0.85	0.883	0.868	0.871	0.935	0.876	0.887	0.898	0.947
0.90	0.861	0.869	0.876	0.939	0.865	0.888	0.897	0.944
0.95	0.861	0.862	0.885	0.924	0.862	0.883	0.893	0.936
1.00	0.840	0.843	0.863	0.921	0.888	0.883	0.889	0.935
1.05	0.864	0.860	0.860	0.908	0.863	0.854	0.872	0.938
1.10	0.840	0.835	0.858	0.896	0.868	0.873	0.844	0.924
1.15	0.832	0.831	0.842	0.882	0.849	0.865	0.860	0.912
1.20	0.831	0.811	0.836	0.903	0.835	0.855	0.847	0.903
1.25	0.832	0.814	0.816	0.862	0.830	0.832	0.840	0.905
1.30	0.779	0.799	0.828	0.868	0.820	0.800	0.846	0.871
1.35	0.790	0.816	0.820	0.878	0.808	0.849	0.831	0.868
1.40	0.806	0.820	0.806	0.856	0.804	0.830	0.848	0.850

Table C.4: The relative frequency of the χ^2 tests based on the asymptotic normality of the maximum likelihood estimator with *p-value*$\in (0.05, 0.95)$ calculated from 1.000 simulations under the `K81`* model. Each data set was a multiple sequence alignment generated on the $\mathcal{T}_{1:2}$ tree with the depth branch length set to the values indicated by the first columns. We present results for a variety of data lengths L and with the distinction as to the positioning of the branch in the tree.

	depth 1				depth 2			
l \ L	300nt	500nt	1.000nt	10.000nt	300nt	500nt	1.000nt	10.000nt
0.01	0.903	0.913	0.939	0.953	0.888	0.891	0.938	0.942
0.05	0.937	0.934	0.946	0.947	0.937	0.942	0.943	0.954
0.10	0.934	0.937	0.953	0.951	0.932	0.942	0.947	0.945
0.15	0.941	0.948	0.942	0.945	0.928	0.939	0.941	0.941
0.20	0.948	0.942	0.956	0.953	0.936	0.946	0.949	0.948
0.25	0.941	0.939	0.934	0.956	0.930	0.950	0.939	0.948
0.30	0.955	0.942	0.958	0.954	0.942	0.938	0.927	0.944
0.35	0.943	0.936	0.955	0.941	0.932	0.949	0.936	0.949
0.40	0.940	0.941	0.946	0.939	0.948	0.931	0.949	0.959
0.45	0.929	0.942	0.940	0.951	0.938	0.943	0.958	0.941
0.50	0.930	0.943	0.944	0.952	0.939	0.947	0.957	0.949
0.55	0.936	0.947	0.930	0.942	0.943	0.938	0.943	0.941
0.60	0.925	0.935	0.939	0.937	0.926	0.935	0.952	0.958
0.65	0.914	0.931	0.927	0.941	0.930	0.942	0.950	0.947
0.70	0.911	0.919	0.928	0.949	0.934	0.926	0.929	0.950
0.75	0.884	0.886	0.926	0.942	0.927	0.925	0.934	0.933
0.80	0.883	0.911	0.926	0.944	0.941	0.918	0.928	0.946
0.85	0.883	0.876	0.915	0.964	0.910	0.939	0.928	0.952
0.90	0.880	0.874	0.902	0.941	0.924	0.922	0.936	0.944
0.95	0.864	0.867	0.895	0.940	0.904	0.935	0.922	0.937
1.00	0.871	0.878	0.885	0.938	0.909	0.920	0.921	0.924
1.05	0.817	0.852	0.895	0.919	0.896	0.919	0.912	0.929
1.10	0.854	0.836	0.884	0.935	0.839	0.905	0.901	0.927
1.15	0.847	0.853	0.871	0.924	0.860	0.879	0.920	0.907
1.20	0.864	0.872	0.860	0.919	0.844	0.883	0.902	0.927
1.25	0.832	0.825	0.854	0.922	0.816	0.862	0.879	0.937
1.30	0.830	0.832	0.844	0.901	0.796	0.847	0.884	0.932
1.35	0.847	0.824	0.864	0.904	0.815	0.860	0.870	0.922
1.40	0.838	0.849	0.827	0.901	0.803	0.818	0.866	0.912

Table C.5: The relative frequency of the χ^2 tests based on the asymptotic normality of the maximum likelihood estimator with *p-value*$\in (0.05, 0.95)$ calculated from 1.000 simulations under the `K81`* model. Each data set was a multiple sequence alignment generated on the $\mathcal{T}_{2:1}$ tree with the depth branch length set to the values indicated by the first columns. We present results for a variety of data lengths L and with the distinction as to the positioning of the branch in the tree.

	depth 1				depth 2			
l \ L	300nt	500nt	1.000nt	10.000nt	300nt	500nt	1.000nt	10.000nt
0.01	0.880	0.891	0.921	0.928	0.918	0.912	0.941	0.950
0.05	0.927	0.916	0.941	0.948	0.936	0.950	0.954	0.942
0.10	0.942	0.942	0.938	0.942	0.955	0.944	0.949	0.950
0.15	0.939	0.938	0.949	0.940	0.951	0.939	0.945	0.953
0.20	0.938	0.938	0.948	0.952	0.950	0.930	0.950	0.935
0.25	0.943	0.933	0.936	0.948	0.955	0.948	0.942	0.949
0.30	0.940	0.945	0.952	0.956	0.950	0.935	0.963	0.956
0.35	0.934	0.938	0.945	0.955	0.947	0.936	0.940	0.955
0.40	0.944	0.947	0.938	0.943	0.933	0.953	0.955	0.940
0.45	0.943	0.949	0.948	0.950	0.937	0.947	0.944	0.941
0.50	0.944	0.951	0.947	0.961	0.939	0.954	0.942	0.951
0.55	0.932	0.932	0.949	0.949	0.941	0.944	0.951	0.954
0.60	0.928	0.940	0.949	0.954	0.944	0.958	0.949	0.946
0.65	0.932	0.943	0.931	0.941	0.951	0.948	0.952	0.949
0.70	0.925	0.946	0.942	0.950	0.939	0.933	0.953	0.962
0.75	0.928	0.932	0.936	0.952	0.937	0.941	0.952	0.943
0.80	0.923	0.919	0.944	0.937	0.940	0.946	0.950	0.953
0.85	0.925	0.935	0.932	0.952	0.941	0.937	0.932	0.947
0.90	0.915	0.926	0.936	0.947	0.938	0.938	0.942	0.951
0.95	0.907	0.930	0.945	0.949	0.928	0.939	0.949	0.939
1.00	0.918	0.924	0.924	0.933	0.931	0.936	0.942	0.954
1.05	0.923	0.907	0.902	0.945	0.907	0.931	0.948	0.946
1.10	0.898	0.910	0.916	0.945	0.913	0.942	0.924	0.950
1.15	0.905	0.903	0.913	0.947	0.902	0.923	0.940	0.950
1.20	0.901	0.899	0.908	0.937	0.885	0.930	0.925	0.966
1.25	0.885	0.897	0.916	0.946	0.889	0.913	0.940	0.945
1.30	0.892	0.906	0.901	0.928	0.896	0.891	0.927	0.952
1.35	0.852	0.895	0.892	0.932	0.883	0.889	0.921	0.954
1.40	0.872	0.863	0.881	0.930	0.883	0.897	0.919	0.947

Table C.6: Variance of $\hat{\xi}, (\times 10^{-2})$, under the JC69* model for $\mathcal{T}^4_{\text{balanced}}$. Branch length varied as listed in the first column. The results are presented for the depth 1 and depth 2 branches and varying multiple sequence alignment lengths.

	depth 1				depth 2			
l \ L	300nt	500nt	1.000nt	10.000nt	300nt	500nt	1.000nt	10.000nt
0.01	0.00037	0.00022	0.00011	0.00001	0.00038	0.00023	0.00011	0.00001
0.05	0.00188	0.00113	0.00056	0.00006	0.00198	0.00119	0.00059	0.00006
0.10	0.00385	0.00231	0.00116	0.00012	0.00427	0.00256	0.00128	0.00013
0.15	0.00602	0.00361	0.00181	0.00018	0.00695	0.00417	0.00209	0.00021
0.20	0.00846	0.00508	0.00254	0.00025	0.01014	0.00608	0.00304	0.00030
0.25	0.01128	0.00677	0.00338	0.00034	0.01395	0.00837	0.00418	0.00042
0.30	0.01461	0.00876	0.00438	0.00044	0.01852	0.01111	0.00556	0.00056
0.35	0.01860	0.01116	0.00558	0.00056	0.02405	0.01443	0.00721	0.00072
0.40	0.02347	0.01408	0.00704	0.00070	0.03075	0.01845	0.00922	0.00092
0.45	0.02946	0.01767	0.00884	0.00088	0.03891	0.02335	0.01167	0.00117
0.50	0.03690	0.02214	0.01107	0.00111	0.04889	0.02934	0.01467	0.00147
0.55	0.04623	0.02774	0.01387	0.00139	0.06116	0.03670	0.01835	0.00183
0.60	0.05800	0.03480	0.01740	0.00174	0.07630	0.04578	0.02289	0.00229
0.65	0.07292	0.04375	0.02188	0.00219	0.09508	0.05705	0.02852	0.00285
0.70	0.09195	0.05517	0.02758	0.00276	0.11847	0.07108	0.03554	0.00355
0.75	0.11630	0.06978	0.03489	0.00349	0.14775	0.08865	0.04432	0.00443
0.80	0.14756	0.08854	0.04427	0.00443	0.18456	0.11073	0.05537	0.00554
0.85	0.18782	0.11269	0.05635	0.00563	0.23103	0.13862	0.06931	0.00693
0.90	0.23978	0.14387	0.07194	0.00719	0.28996	0.17397	0.08699	0.00870
0.95	0.30700	0.18420	0.09210	0.00921	0.36495	0.21897	0.10948	0.01095
1.00	0.39408	0.23645	0.11822	0.01182	0.46071	0.27643	0.13821	0.01382
1.05	0.50707	0.30424	0.15212	0.01521	0.58339	0.35004	0.17502	0.01750
1.10	0.65382	0.39229	0.19615	0.01961	0.74098	0.44459	0.22229	0.02223
1.15	0.84462	0.50677	0.25339	0.02534	0.94389	0.56633	0.28317	0.02832
1.20	1.09288	0.65573	0.32786	0.03279	1.20570	0.72342	0.36171	0.03617
1.25	1.41610	0.84966	0.42483	0.04248	1.54412	0.92647	0.46324	0.04632
1.30	1.83715	1.10229	0.55114	0.05511	1.98223	1.18934	0.59467	0.05947
1.35	2.38585	1.43151	0.71576	0.07158	2.55012	1.53007	0.76504	0.07650
1.40	3.10117	1.86070	0.93035	0.09304	3.28704	1.97222	0.98611	0.09861

Table C.7: Variance of $\hat{\xi}, (\times 10^{-2})$, under the JC69* model for $\mathcal{T}_{1:2}$; see Tab. C.6 for a more detailed description

	depth 1				depth 2			
l \ L	300nt	500nt	1.000nt	10.000nt	300nt	500nt	1.000nt	10.000nt
0.01	0.00037	0.00022	0.00011	0.00001	0.00019	0.00011	0.00006	0.00001
0.05	0.00182	0.00109	0.00055	0.00005	0.00109	0.00065	0.00033	0.00003
0.10	0.00364	0.00218	0.00109	0.00011	0.00256	0.00154	0.00077	0.00008
0.15	0.00549	0.00329	0.00165	0.00016	0.00454	0.00272	0.00136	0.00014
0.20	0.00743	0.00446	0.00223	0.00022	0.00715	0.00429	0.00215	0.00021
0.25	0.00951	0.00571	0.00285	0.00029	0.01057	0.00634	0.00317	0.00032
0.30	0.01179	0.00707	0.00354	0.00035	0.01501	0.00901	0.00450	0.00045
0.35	0.01432	0.00859	0.00430	0.00043	0.02071	0.01243	0.00621	0.00062
0.40	0.01717	0.01030	0.00515	0.00051	0.02801	0.01681	0.00840	0.00084
0.45	0.02041	0.01224	0.00612	0.00061	0.03730	0.02238	0.01119	0.00112
0.50	0.02413	0.01448	0.00724	0.00072	0.04908	0.02945	0.01472	0.00147
0.55	0.02845	0.01707	0.00853	0.00085	0.06401	0.03840	0.01920	0.00192
0.60	0.03349	0.02010	0.01005	0.00100	0.08288	0.04973	0.02486	0.00249
0.65	0.03942	0.02365	0.01183	0.00118	0.10673	0.06404	0.03202	0.00320
0.70	0.04642	0.02785	0.01393	0.00139	0.13686	0.08212	0.04106	0.00411
0.75	0.05473	0.03284	0.01642	0.00164	0.17495	0.10497	0.05248	0.00525
0.80	0.06462	0.03877	0.01939	0.00194	0.22312	0.13387	0.06694	0.00669
0.85	0.07644	0.04586	0.02293	0.00229	0.28411	0.17047	0.08523	0.00852
0.90	0.09060	0.05436	0.02718	0.00272	0.36141	0.21685	0.10842	0.01084
0.95	0.10760	0.06456	0.03228	0.00323	0.45952	0.27571	0.13786	0.01379
1.00	0.12806	0.07684	0.03842	0.00384	0.58421	0.35053	0.17526	0.01753
1.05	0.15272	0.09163	0.04582	0.00458	0.74292	0.44575	0.22288	0.02229
1.10	0.18249	0.10949	0.05475	0.00547	0.94523	0.56714	0.28357	0.02836
1.15	0.21846	0.13108	0.06554	0.00655	1.20349	0.72209	0.36105	0.03610
1.20	0.26198	0.15719	0.07859	0.00786	1.53366	0.92020	0.46010	0.04601
1.25	0.31467	0.18880	0.09440	0.00944	1.95636	1.17381	0.58691	0.05869
1.30	0.37851	0.22711	0.11355	0.01136	2.49821	1.49893	0.74946	0.07495
1.35	0.45591	0.27355	0.13677	0.01368	3.19368	1.91621	0.95810	0.09581
1.40	0.54982	0.32989	0.16495	0.01649	4.08736	2.45242	1.22621	0.12262

Table C.8: Variance of $\hat{\xi}, (\times 10^{-3})$, under the `JC69`* model for $\mathcal{T}_{2:1}$; see Tab. C.6 for a more detailed description.

	depth 1				depth 2			
l \ L	300nt	500nt	1.000nt	10.000nt	300nt	500nt	1.000nt	10.000nt
0.01	0.00186	0.00112	0.00056	0.00006	0.00368	0.00221	0.00110	0.00011
0.05	0.00956	0.00573	0.00287	0.00029	0.01805	0.01083	0.00541	0.00054
0.10	0.01985	0.01191	0.00596	0.00060	0.03533	0.02120	0.01060	0.00106
0.15	0.03113	0.01868	0.00934	0.00093	0.05212	0.03127	0.01564	0.00156
0.20	0.04367	0.02620	0.01310	0.00131	0.06862	0.04117	0.02059	0.00206
0.25	0.05779	0.03468	0.01734	0.00173	0.08505	0.05103	0.02551	0.00255
0.30	0.07388	0.04433	0.02216	0.00222	0.10158	0.06095	0.03047	0.00305
0.35	0.09238	0.05543	0.02771	0.00277	0.11841	0.07105	0.03552	0.00355
0.40	0.11385	0.06831	0.03416	0.00342	0.13571	0.08142	0.04071	0.00407
0.45	0.13895	0.08337	0.04168	0.00417	0.15366	0.09219	0.04610	0.00461
0.50	0.16845	0.10107	0.05054	0.00505	0.17245	0.10347	0.05173	0.00517
0.55	0.20333	0.12200	0.06100	0.00610	0.19227	0.11536	0.05768	0.00577
0.60	0.24474	0.14685	0.07342	0.00734	0.21335	0.12801	0.06400	0.00640
0.65	0.29409	0.17646	0.08823	0.00882	0.23590	0.14154	0.07077	0.00708
0.70	0.35309	0.21186	0.10593	0.01059	0.26019	0.15611	0.07806	0.00781
0.75	0.42382	0.25429	0.12715	0.01271	0.28648	0.17189	0.08594	0.00859
0.80	0.50881	0.30529	0.15264	0.01526	0.31510	0.18906	0.09453	0.00945
0.85	0.61114	0.36668	0.18334	0.01833	0.34638	0.20783	0.10391	0.01039
0.90	0.73456	0.44074	0.22037	0.02204	0.38073	0.22844	0.11422	0.01142
0.95	0.88365	0.53019	0.26510	0.02651	0.41859	0.25115	0.12558	0.01256
1.00	1.06398	0.63839	0.31919	0.03192	0.46046	0.27627	0.13814	0.01381
1.05	1.28236	0.76941	0.38471	0.03847	0.50691	0.30415	0.15207	0.01521
1.10	1.54708	0.92825	0.46412	0.04641	0.55860	0.33516	0.16758	0.01676
1.15	1.86827	1.12096	0.56048	0.05605	0.61625	0.36975	0.18488	0.01849
1.20	2.25831	1.35498	0.67749	0.06775	0.68073	0.40844	0.20422	0.02042
1.25	2.73227	1.63936	0.81968	0.08197	0.75299	0.45179	0.22590	0.02259
1.30	3.30860	1.98516	0.99258	0.09926	0.83412	0.50047	0.25024	0.02502
1.35	4.00980	2.40588	1.20294	0.12029	0.92539	0.55523	0.27762	0.02776
1.40	4.86337	2.91802	1.45901	0.14590	1.02821	0.61692	0.30846	0.03085

Table C.9: Mean of the variances of $\hat{\xi}_b(\times 10^{-1})$ for 1.000 samples generated under `K81`* model for $\mathcal{T}^4_{\text{balanced}}$. Branch length of different sets are given in the first column. Depth 1 branches refer to the branches leading to the leves of the tree, depth 2 to the interior ones. Having confirmed that both the sets of inner and external branches give virtually same results, here we depict the results for a selected branch of each of the sets.

	depth 1				depth 2			
l \ L	300nt	500nt	1.000nt	10.000nt	300nt	500nt	1.000nt	10.000nt
0.01	0.00018	0.00014	0.00011	0.00008	0.00019	0.00014	0.00011	0.00008
0.05	0.00229	0.00206	0.00194	0.00168	0.00231	0.00199	0.00179	0.00168
0.10	0.00728	0.00690	0.00650	0.00619	0.00750	0.00728	0.00657	0.00627
0.15	0.01476	0.01367	0.01341	0.01280	0.01589	0.01414	0.01330	0.01274
0.20	0.02424	0.02293	0.02155	0.02158	0.02484	0.02291	0.02172	0.02155
0.25	0.03438	0.03237	0.03060	0.02972	0.03589	0.03329	0.03187	0.03152
0.30	0.04358	0.04302	0.04151	0.04021	0.04760	0.04402	0.04185	0.04180
0.35	0.05783	0.05270	0.05124	0.04880	0.06181	0.05457	0.05414	0.04962
0.40	0.06954	0.06781	0.06116	0.06227	0.07457	0.06954	0.06232	0.06212
0.45	0.08237	0.07920	0.07715	0.07277	0.08813	0.08388	0.07388	0.07210
0.50	0.09555	0.08855	0.08448	0.07853	0.10567	0.09428	0.08703	0.08699
0.55	0.10950	0.10187	0.09681	0.08900	0.12294	0.11033	0.10166	0.09291
0.60	0.12841	0.11407	0.10603	0.10280	0.14349	0.13227	0.11281	0.10348
0.65	0.14392	0.12495	0.11865	0.10817	0.17056	0.14612	0.12505	0.10895
0.70	0.16076	0.14408	0.12603	0.11740	0.19245	0.16216	0.13681	0.11853
0.75	0.19569	0.15809	0.14051	0.12159	0.22870	0.18760	0.15746	0.12817
0.80	0.21708	0.18257	0.15992	0.13510	0.26844	0.21832	0.17131	0.13211
0.85	0.25384	0.20428	0.17306	0.14160	0.31971	0.24989	0.18872	0.14631
0.90	0.29724	0.23078	0.19043	0.14824	0.39092	0.29057	0.21395	0.14772
0.95	0.35940	0.27887	0.20979	0.15236	0.48425	0.33459	0.24870	0.16324
1.00	0.45058	0.33457	0.24576	0.16853	0.55092	0.41060	0.27174	0.17126
1.05	0.55825	0.41069	0.25767	0.16448	0.72024	0.47179	0.32817	0.17567
1.10	0.71063	0.49548	0.32344	0.18476	0.88286	0.60233	0.37592	0.19163
1.15	0.90301	0.58253	0.40023	0.19254	1.06889	0.71626	0.45984	0.19183
1.20	1.14271	0.88000	0.43978	0.20121	1.35532	0.88789	0.54025	0.20816
1.25	1.57133	0.97347	0.56155	0.21152	1.72189	1.11298	0.65992	0.22020
1.30	2.17904	1.24180	0.76257	0.23530	2.25637	1.38997	0.74390	0.23707
1.35	3.01652	1.87197	0.94136	0.28747	2.66188	1.76640	0.90896	0.25833
1.40	3.41393	2.13294	1.12709	0.28467	3.46262	2.00757	1.13683	0.28258

Table C.10: Mean of the variances of $\hat{\xi}_b(\times 10^{-1})$ for 1.000 samples generated under the K81* model for $\mathcal{T}_{1:2}$; see Tab. C.9 for a more detailed description.

	depth 1				depth 2			
l \ L	300nt	500nt	1.000nt	10.000nt	300nt	500nt	1.000nt	10.000nt
0.01	0.00018	0.00014	0.00011	0.00008	0.00008	0.00005	0.00004	0.00002
0.05	0.00232	0.00207	0.00192	0.00164	0.00076	0.00063	0.00055	0.00046
0.10	0.00745	0.00702	0.00651	0.00613	0.00249	0.00225	0.00195	0.00172
0.15	0.01406	0.01410	0.01326	0.01281	0.00513	0.00458	0.00410	0.00362
0.20	0.02347	0.02277	0.02105	0.02114	0.00895	0.00793	0.00681	0.00635
0.25	0.03266	0.03184	0.03117	0.02977	0.01291	0.01168	0.01046	0.00954
0.30	0.04494	0.04263	0.04071	0.03980	0.01903	0.01641	0.01403	0.01286
0.35	0.05436	0.05194	0.05036	0.05002	0.02480	0.02183	0.01961	0.01687
0.40	0.06504	0.06199	0.06241	0.05970	0.03337	0.02765	0.02490	0.02096
0.45	0.07864	0.07736	0.07249	0.07178	0.04215	0.03538	0.03073	0.02576
0.50	0.09083	0.08504	0.08386	0.07958	0.05360	0.04454	0.03703	0.03150
0.55	0.10274	0.09519	0.09418	0.09036	0.06636	0.05363	0.04496	0.03413
0.60	0.11975	0.10880	0.10321	0.10020	0.08375	0.06578	0.05299	0.04195
0.65	0.12681	0.11813	0.11548	0.10676	0.10353	0.08069	0.06109	0.04788
0.70	0.14201	0.12794	0.12355	0.11455	0.12777	0.09733	0.07416	0.05101
0.75	0.15499	0.14253	0.13283	0.12752	0.16369	0.12087	0.08929	0.05698
0.80	0.16729	0.15023	0.14369	0.13359	0.19963	0.14677	0.10226	0.06330
0.85	0.18138	0.15917	0.15065	0.13625	0.25722	0.17464	0.11994	0.07099
0.90	0.19610	0.17898	0.15808	0.14101	0.30956	0.22228	0.14488	0.07918
0.95	0.21081	0.19149	0.17013	0.14662	0.39692	0.26992	0.16843	0.08435
1.00	0.23890	0.20060	0.18327	0.15518	0.49922	0.33282	0.20669	0.09534
1.05	0.26122	0.21839	0.19455	0.16155	0.63976	0.41998	0.25218	0.10356
1.10	0.28720	0.24293	0.19973	0.16865	0.81820	0.52855	0.29843	0.11144
1.15	0.31557	0.26544	0.21485	0.16769	1.04257	0.63603	0.38582	0.12646
1.20	0.38017	0.29657	0.22980	0.17678	1.22670	0.78909	0.45575	0.13134
1.25	0.42516	0.31489	0.24917	0.18657	1.65925	1.04029	0.56345	0.15076
1.30	0.53650	0.37458	0.28482	0.18367	1.98764	1.25034	0.66879	0.16397
1.35	0.54247	0.41055	0.31067	0.19938	2.59005	1.57265	0.85285	0.18268
1.40	0.73088	0.49620	0.33082	0.20203	3.21930	2.13819	1.08728	0.21025

Table C.11: Mean of the variances of $\hat{\xi}_b(\times 10^{-2})$ for 1.000 samples generated under K81* model for $\mathcal{T}_{2:1}$; see Tab. C.9 for a more detailed description.

	depth 1				depth 2			
l \ L	300nt	500nt	1.000nt	10.000nt	300nt	500nt	1.000nt	10.000nt
0.01	0.00074	0.00051	0.00036	0.00020	0.00186	0.00141	0.00107	0.00074
0.05	0.00749	0.00626	0.00546	0.00432	0.02185	0.01984	0.01894	0.01748
0.10	0.02360	0.02091	0.01881	0.01660	0.07245	0.07142	0.06719	0.06160
0.15	0.04561	0.04407	0.04049	0.03681	0.14638	0.14216	0.13142	0.12838
0.20	0.07807	0.07232	0.06728	0.06232	0.24006	0.22111	0.21929	0.20567
0.25	0.11301	0.10300	0.10160	0.09427	0.32681	0.30199	0.30862	0.30129
0.30	0.15460	0.14555	0.13833	0.12717	0.43849	0.42439	0.42265	0.39871
0.35	0.19420	0.18506	0.17538	0.17661	0.53409	0.53822	0.53512	0.50094
0.40	0.24245	0.22925	0.21930	0.20893	0.66888	0.63529	0.61988	0.60905
0.45	0.30137	0.29153	0.26680	0.26364	0.76290	0.76709	0.73038	0.70568
0.50	0.35211	0.33463	0.32191	0.30035	0.89706	0.87194	0.85310	0.81615
0.55	0.42586	0.38561	0.37222	0.33929	0.98352	0.94092	0.93898	0.91090
0.60	0.48491	0.46815	0.43570	0.39439	1.06651	1.06439	1.00178	1.00523
0.65	0.57046	0.50243	0.47887	0.46605	1.16397	1.10825	1.08809	1.09697
0.70	0.65148	0.57090	0.55549	0.50230	1.30672	1.23431	1.20744	1.16367
0.75	0.72819	0.64025	0.59429	0.54764	1.35492	1.31490	1.28772	1.22821
0.80	0.81837	0.72212	0.67485	0.61699	1.47545	1.44807	1.34588	1.28428
0.85	0.94583	0.81262	0.74004	0.69612	1.55086	1.49499	1.41472	1.41714
0.90	1.00244	0.91354	0.79548	0.72438	1.65548	1.57179	1.47219	1.46580
0.95	1.12750	0.97410	0.85839	0.73565	1.71613	1.67795	1.58191	1.48356
1.00	1.26722	1.09787	0.96212	0.80407	1.79343	1.71844	1.61721	1.59481
1.05	1.45500	1.24422	1.04389	0.88240	1.88094	1.83908	1.68888	1.56427
1.10	1.60333	1.34616	1.14877	0.94870	2.03696	1.85232	1.76646	1.65058
1.15	1.81467	1.44897	1.19329	0.96765	2.02791	1.89747	1.84157	1.68069
1.20	2.09509	1.65216	1.33334	0.97023	2.22272	2.08779	1.87387	1.77728
1.25	2.43942	1.87843	1.45003	1.07832	2.33224	2.11088	1.91557	1.76277
1.30	2.76513	2.19626	1.58806	1.13166	2.43266	2.18741	1.97962	1.81825
1.35	3.23513	2.34197	1.74883	1.18653	2.51333	2.18871	2.03981	1.82161
1.40	3.83282	2.73735	1.94607	1.30914	2.69707	2.27705	2.04106	1.91677

Table C.12: The relative frequency of the χ^2 tests based on the asymptotic normality of the maximum likelihood estimator with *p-value* $\in (0.05, 0.95)$ (*left*) and the mean of the variances of ξ_b (*right*) calculated from 1.000 simulations under the K81* model. Each data set was a multiple sequence alignment generated on the $\mathcal{T}^6_{\text{balanced}}$ tree with branch lengths set to 0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3 and 1.4. The data lengths L were taken to be 1.000nt to 10.000nt and the distnction was made as to the positioning of the branches in the tree. The results refer to a chosen branch from the sets of internal and external ones.

	depth 1		depth 2	
l \ L	1.000nt	10.000nt	1.000nt	10.000nt
0.01	0.876	0.937	0.917	0.945
0.10	0.935	0.942	0.937	0.943
0.30	0.961	0.953	0.938	0.956
0.50	0.927	0.951	0.936	0.947
0.70	0.886	0.941	0.889	0.924
0.90	0.849	0.895	0.887	0.902
1.10	0.819	0.854	0.825	0.884
1.30	0.780	0.813	0.733	0.862
1.40	0.753	0.815	0.719	0.833

depth 1		depth 2	
1.000nt	10.000nt	1.000nt	10.000nt
0.00001	0.00001	0.00001	0.00001
0.00065	0.00063	0.00067	0.00061
0.00408	0.00406	0.00434	0.00409
0.00894	0.00798	0.00972	0.00845
0.01511	0.01213	0.02067	0.01327
0.03154	0.01601	0.06118	0.01936
0.10086	0.02588	0.25357	0.03779
0.45042	0.06577	1.15435	0.12553
1.16442	0.15901	2.34536	0.26316

Bibliography

Abascal, F., Zardoya, R., and Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, 21:2104–2105, 2005. Applications note.

Abbott, J. and Bigatti, A. *CoCoALib: A C++ library for Computations in Commutative Algebra... and Beyond*, volume 6327 of *Lecture Notes in Comp.Sci.*, pages 73–76. Springer, 2010. Invited talk.

Abbott, J., Bigatti, A., Caboara, M., and Robbiano, L. Cocoa: Computations in commutative algebra. *SIGSAM Communications in Computer Algebra*, 2007. also presented as Software Demo of the ISSAC07 Conference.

Adachi, J. and Hasegawa, M. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.*, 42(4):459–68, 1996.

Aheng, D., Franksich, A., Baertsch, R., Kapranov, F., Reymond, A., Choo, S. W., Lu, Y., Denoeud, F., Antonarakis, S. E., Snyder, M., and et.al. Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. *Genome Res.*, 17:839–851, 2007.

Allman, E. S. and Rhodes, J. A. Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical Biosciences*, 186(2):113–144, 2003.

Allman, E. S. and Rhodes, J. A. Quartets and parameter recovery for the general Markov model of sequence mutation. *AMRX Applied Mathematics Research Express*, 2004(4):107–131, 2004a. ISSN 1687-1200.

Allman, E. S. and Rhodes, J. A. *Mathematical Models in Biology.* Cambridge University Press, January 2004b. ISBN 0-521-52586-1.

Allman, E. S. and Rhodes, J. A. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *Journal of Computational Biology*, 13:1101–1113, 2006a.

Allman, E. S. and Rhodes, J. A. Phylogenetic invariants for stationary base composition. *Journal of Symbolic Computation*, 41(2):138 – 150, 2006b. Computational Algebraic Statistics.

Allman, E. S. and Rhodes, J. A. *Reconstructing Evolution*, chapter Phylogenetic invariants. Gascuel, O and Steel, M A (ed.); Oxford University Press, 2007.

Allman, E. S. and Rhodes, J. A. Phylogenetic ideals and varieties for the general Markov model. *Adv. in Appl. Math.*, 40:127–148, 2008a.

Allman, E. S. and Rhodes, J. A. Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites. *Mathematical Biosciences*, 211:18–33, 2008b.

Allman, E. S., Petrovic, S., Rhodes, J. A., and Sullivant, S. Identifiability of two-tree mixtures for group-based models. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 2010.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *J. Mol, Biol*, 3:403—410, October 1990.

Ané, C., Burleigh, J. G., McMahon, M. M., and Sanderson, M. J. Covarion structure in plastid genome evolution: A new statistical test. *Mol. Biol. Evol.*, 22(4):914–924, 2005.

Balakirev, E. S. and Ayala, F. J. Pseudogenes: are they "junk" or functional DNA? *Annu Rev Genet.*, 37:123–51, 2003.

Balakirev, E. and Ayala, F. Pseudogenes: structure conservation, expression, and functions. *In Zhurnal Obshchei Biologii*, 65(4):306–321, 2004.

Barrette, I. H., McKenna, S., Taylor, D. R., and Forsdyke, D. R. Introns resolve the conflict between base order-dependent stem-loop potential and the encoding of RNA or protein: further evidence from overlapping genes. *Gene*, 270(1-2):181 – 189, 2001.

Barry, D. and Hartigan, J. A. Asynchronous distance between homologous DNA sequences. *Biometrics*, 43(2):261–276, 1987. ISSN 0006-341X.

Bensasson, D., Petrov, D., Zhang, D., Hartl, D., and Hewitt, G. Genomic gigantism: DNA loss is slow in mountain grasshoppers. *Mol. Biol. and Evol*, 18:246–253, 2001.

Bray, N. and Pachter, L. MAVID multiple alignment server. *Nucleic Acids Res*, 31: 3525–6, 2003.

Bruno, W. and Halpern, A. Topological bias and inconsistency of maximum likelihood using wrong models. *Mol. Biol. Evol.*, 16(4):564–566, 1999.

Burnham, K. P. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304, November 2004. ISSN 0049-1241.

Carlini, D. B. and Genut, J. E. Synonymous *SNP*s provide evidence for selective constraint on human exonic splicing enhancers. *J. Mol. Biol.*, 62(1):89–98, 2006.

Casanellas, M. and Fernández-Sánchez, J. Performance of a New Invariants Method on Homogeneous and Nonhomogeneous Quartet Trees. *Mol. Biol. Evol.*, 24(1):288–293, 2007.

Casanellas, M. and Fernández-Sánchez, J. Geometry of the Kimura 3-parameter model. *Advances in Applied Mathematics*, 41(3):265–292, 2008.

Casanellas, M. and Fernández-Sánchez, J. Relevant phylogenetic invariants of evolutionary models. *J. Mathématiques Pures et Appliquées*, 96(3):207–229, 2011.

Casanellas, M. and Kedzierska, A. M. Generating Markov evolutionary matrices for a given branch length. submitted, 2011. URL `http://arxiv.org/abs/1112.3529`.

Casanellas, M. and Sullivant, S. The strand symmetric model. In Pachter, L. and Sturmfels, B., editors, *Algebraic Statistics for computational biology*, chapter 16. Cambridge University Press, 2005.

Casanellas, M., Fernández-Sánchez, J., and Kedzierska, A. M. The space of phylogenetic mixtures of equivariant models. submitted to the special issue of Alg. for Mol. Biol., 2011. URL `http://arxiv.org/abs/1110.0920`.

Cavender, J. and Felsenstein, J. Invariants of phylogenies in a simple case with discrete states. *J. Classification*, 4:57–71, 1987.

Chai, J. and Housworth, E. On Rogers's proof of identifiability for the GTR + Gamma + I model. *Syst. Biol.*, 2011.

Chamary, J. V., Parmley, J. L., and Hurst, L. D. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*, 7:98–108, 2006.

Chang, J. T. Inconsistency of Evolutionary Tree Topology Reconstruction Methods When Substitution Rates Vary Across Characters, 1994.

Chang, J. T. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Bios.*, 137(1):51–73, 1996. ISSN 0025-5564.

Chang, K., Pearson, K., and Zhang, T. Perron-Frobenius theorem for nonnegative tensors. *Commun. Math. Sci.*, 6:507–520, 2008.

Chen, K. and Rajewsky, N. Deep conservation of microRNA-target relationships and 3'UTR motifs in vertebrates, flies, and nematodes. *Cold Spring Harb Symp Quant Biol.*, pages 149–56, 2006.

Churbanov, A., Rogozin, I. B., Babenko, V. N., Ali, H., and Koonin, E. V. Evolutionary conservation suggests a regulatory function of AUG triplets in 5'UTRs of eukaryotic genes. *Nucleic Acids Res.*, 33(17):5512–5520, 2005.

Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, N., and Pollard, D. A. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167):203–18, November 2007.

Cohen, J. E. Mathematics Is Biology's Next Microscope, Only Better; Biology Is Mathematics' Next Physics, Only Better. *PLoS Biol*, 2(12):e439, 12 2004.

Cox, D., Little, J., and O'Shea, D. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer, 3rd edition, 2007.

Crick and Watson. Molecular structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171:737—-738, 1953.

Culver, W. J. On the existence and uniqueness of the real logarithm of a matrix. *Proc. Amer. Math. Soc.*, 17:1146–1151, 1966.

Darwin, C. *On the Origin of Species*. Watts and Co., London, 1929. 6th edition.

Dayhoff, M., Schwartz, R., and Orcutt, B. A model of evolutionary change in proteins. *In Dayhoff, M.O. (Ed.). Atlas of Protein Sequence and Structure , W-n, DC Nat. Biom. Res. F-n*, page 345, 1978.

Dempster, A., Laird, N., and Rubin, D. Maximum Likelihood from incomplete data. *J. Royal. Stat. Soc.*, 39(1):1–38, 1977.

Diaconis, P. and Sturmfels, B. Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26(1):363–397, March 1998. ISSN 0090-5364.

Draisma, J. and Kuttler, J. On the ideals of equivariants tree models. *Mathematische Annalen*, 344:619–644, 2009.

Drton, M. Discrete chain graph models. *Bernoulli*, 15:736–753, 2008.

Drton, M. Likelihood ratio tests and singularities. *Annals of Statistics*, 37:979–1012, 2009.

Drton, M. and Foygel, R. Extended Bayesian information criteria for Gaussian graphical models. *Advances in Neural Inf. Proc. Sys.*, 23:2020–2028, 2008.

Drton, M. and Richardson, T. S. Graphical methods for efficient likelihood inference in Gaussian covariance models. *Journal of Machine Learning Research*, 9:893–914, May 2008.

Drton, M. and Sullivant, S. Algebraic statistical models. *Statistica Sinica*, 17:1273–1297, 2007.

Drton, M., Chaudhuri, S., and Richardson, T. S. Estimation of a covariance matrix with zeros. *Biometrika*, 94:199–216, 2007.

Drton, M., Sturmfels, B., and Sullivant, S. *Lectures on Algebraic Statistics (Oberwolfach Seminars)*. Birkhäuser Basel, 2008. ISBN 3764389044.

Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32:1792–97, 2004.

Efron, B. and Hinkley, D. V. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, 65:457–483, 1978.

Eriksson, N. Tree construction using singular value decomposition. In Pachter, L. and Sturmfels, B., editors, *Algebraic Statistics for computational biology*, chapter 19, pages 347–358. Cambridge University Press, 2005.

Fairbrother, W., Yeh, R., Sharp, P., and Burge, C. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297:1007–1013, 2002.

Fairbrother, W. G., Holste, D., Burge, C. B., and Sharp, P. A. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS biology*, 2 (9) E268, 2004.

Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.5c., 1993.

Felsenstein, J. *Inferring Phylogenies.* Sinauer Associates, Inc., 2003.

Fienberg, S. E. *The Analysis of Cross-classified Categorical Data.* 2nd Edition. M.I.T. Press, Cambridge, 1980.

Foster, P. Modeling compositional heterogeneity. *Systematic Biology*, 53(3):485–495, 2004.

Fu, Y.-X. and Li, W.-H. Construction of linear invariants in phylogenetic inference. *Mathematical Biosciences*, 109(2):201 – 228, 1992a. ISSN 0025-5564.

Fu, Y. and Li, W. Construction of linear invariants in phylogenetic inference. *Mathematical Biosciences*, 109(2):201–228, 1992b.

Fujita, P., Rhead, B., Zweig, A., Hinrichs, A., Karolchik, D., Cline, M., Goldman, M., Barber, G., Clawson, H., Coelho, A., and et.al. The UCSC Genome Browser database: update 2011. *Nucleic Acids Research*, 39:1–7, 2010.

Galtier, N. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.*, 18(5):866–873, 2001a.

Galtier, N. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. and Evol.*, 18(5):866–873, 2001b.

Garcia-Puente, L. D. Algebraic statistics in model selection. *Proceedings of the 20th Conf. on Uncertainty in Art. Intel.*, pages 177–184, 2004.

Garcia-Puente, L., Stillman, M., and Sturmfels, B. Algebraic geometry for Bayesian networks. *Journal of Symb. Comp.*, 39:331–355, 2005.

Gascuel, O. and Guindon, S. *Reconstructing Evolution: new mathematical and computational advances.*, chapter Modelling the variability of evolutionary processes. Oxford University Press, 2007. Gascuel, O and Steel, M A (ed.).

Gaucher, E. A., Miyamoto, M. M., and Benner, S. A. Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. *Proc. Natl. Acad. Sci. USA*, 98(2):548–552, 2001.

Gibilisco, P., Riccomagno, E., Rogantin, M., and Wynn, H. P. *Algebraic and Geometric Methods in Statistics.* Cambridge University Press, 2009. 1st Edition.

Goldman, N. and Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol*, 11(5):725–736, 1994.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., and Ast, G. Comparative analysis identifies exonic splicing regulatory sequences - the complex definition of enhancers and silencers. *Mol. Cell*, 22:769–781, 2006.

Grayson, D. R. and Stillman, M. E. Macaulay2, a software system for research in algebraic geometry. Available at `http://www.math.uiuc.edu/Macaulay2/`, 2009.

Greuel, G. M., Pfister, G., and Schönemann, H. *Symbolic computation and automated reasoning, The Calculemus-2000 Symposium*, chapter SINGULAR 3.0 — A computer algebra system for polynomial computations, pages 227–233. Kerber, M and Kohlhase, M (ed.), A. K. Peters, Ltd., 2001. ISBN 1-56881-145-4.

Greuel, G., Pfister, G., and Schoenemann, H. Singular: A computer algebra system for polynomial computations. Available at `http://www.singular.uni-kl.de/`, 2003.

Guindon, S. and Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52:696–704, 2003.

Guindon, S., Rodrigo, A. G., Dyer, K. A., and Huelsenbeck, J. P. Modelling the site-specific variation of selection patterns along lineages. *Proc. Natl. Acad. Sci. USA*, 101 (35):12957–12962, 2004.

Harris, J. *Algebraic geometry. A first course*, volume 133 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1992. ISBN 0-387-97716-3.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C., Chrast, J., Lagarde, J., Gilbert, J., Storey, R., Swarbreck, D., and et. al. GENCODE: producing a reference annotation for ENCODE. *Genome Biol*, 7 (Suppl 1), 2006. S4.1–S4.9.

Hartley, H. Maximum likelihood estimation from incomplete data. *Biometrics*, 14: 174–194, 1958.

Hartshorne, R. *Algebraic Geometry.* Springer, 1977.

Hasegawa, M., Kishino, H., and Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Biol.*, 22(2):160–174, 1985.

He, K. and Meeden, G. Selecting the number of bins in a histogram: A decision theoretic approach. *Journal of Statistical Planning and Inference*, 61:59–76, 1997.

Hillis, D., Huelsenbeck, J., and Swofford, D. Hobglobin of phylogenetics? *Nature*, 369(6479):363–364, 1994.

Ho, S. and Jermiin, L. Tracing the decay of the historical signal in biological sequence data. *Syst. Biol.*, 53(4):623–637, 2004.

Hubbard, T., Aken, B., Ayling, S., and Ballester, B. Ensembl 2009. *Nucleic Acids Research*, 34:D556-61, Jan 2009.

Huelsenbeck, J. P. Testing a covariotide model of DNA substitution. *Mol. Biol. Evol.*, 19(5):698–707, 2002.

Huelsenbeck, J. Performance of phylogenetic methods in simulation. *Systematic Biology*, 44(1):17–48, 1995. ISSN 10635157.

Huelsenbeck, J., Larget, B., and Alfaro, M. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol. Biol. and Evol.*, 21(6):1123–1133, 2004.

Hurst, L. D. Preliminary assessment of the impact of microRNA-mediated regulation on coding sequence evolution in mammals. *J. Mol. Biol.*, 63(2):174–182, 2006.

Husmeier, D., Dybowski, R., and Roberts, S. *Probabilistic Modeling in Bioinformatics and Medical Informatics Advanced Information and Knowledge Processing*. Springer Verlag, New York, 2005.

Irimia, M., Rukov, J. L., Penny, D., Garcia-Fernandez, J., Vinther, J., and Roy, S. W. Widespread Evolutionary Conservation of Alternatively Spliced Exons in Caenorhabditis. *Mol. Biol. and Evol.*, 25(2):375–382, 2008.

Jakse, J., Meyer, J., Suzuki, G., McCallum, J., Cheung, F., Town, C., and Havey, M. Pilot sequencing of onion genomic DNA reveals fragments of transposable elements, low gene densities, and significant gene enrichment after methyl filtration. *Mol. Genetics and Genomics*, 280(4):287–292, 2008.

Jayaswal, V., Jermiin, L. S., Poladian, L., and Robinson, J. Two stationary nonhomogeneous Markov models of nucleotide sequence evolution. *Systematic Biology*, 60(1): 74–86, 2011.

Jermiin, L. S., Ho, S. Y., Ababneh, F., Robinson, J., and Larkum, A. W. Hetero: a program to simulate the evolution of DNA on a four-taxon tree. *Applied Bioinformatics*, 2(3):159–163, 2003.

Jermiin, L. S., Ho, S. Y. W., Ababneh, F., Robinson, J., and Larkum, A. W. The Biasing Effect of Compositional Heterogeneity on Phylogenetic Estimates May be Underestimated. *Syst. Biol.*, 53(4):638–643, 2004.

Jin, L. and Nei, M. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.*, 7:82–102, 1990.

Jones, D. T., Taylor, W. R., and Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comp. Appl. Biosci.*, 8275, 1992.

Jukes, T. and Cantor, C. Evolution of protein molecules. *In Mammalian Protein Metabolism*, pages 21–132, 1969a.

Jukes, T. and Cantor, C. *Evolution of protein molecules*, volume 3 of *Mammalian protein metabolism*, chapter III, pages 21–132. Academic Press, 1969b.

Kamal, M., Xie, X., and Lander, E. A large family of ancient repeat elements in the human genome is under strong selection. *PNAS*, 103(8):2740–2745, 2006.

Ke, S., Zhang, X. H. F., and Chasin, L. A. Positive selection acting on splicing motifs reflects compensatory evolution. *Genome Research*, 18:533–543, Jan 2008.

Kedzierska, A. M., Drton, M., Guigó, R., and Casanellas, M. `SPIn`: model selection for phylogenetic mixtures via linear invariants. *Mol. Biol. Evol.*, 29(3):929–937, 2012.

Kendall, W. Computer algebra in probability and statistics. *Statistica Neerlandica*, 47: 9–25, 1993.

Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Biol.*, 16(2):111–120, 1980. ISSN 00222844. doi: 10.1007/BF01731581.

Kimura, M. Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences of the USA*, 78(1):454–458, 1981.

Kolaczkowski, B. and Thornton, J. W. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, 431(7011):980–984, 2004.

Lake, J. A. A rate-independent technique for analysis of nucleaic acid sequences: evolutionary parsimony. *Mol. Biol. Evol.*, 4:167–191, 1987.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., and et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

Lockhart, P., Larkum, A., Steel, M., Waddell, P., and Penny, D. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. *Proc. Natl. Acad. Sci. USA*, 93(5):1930–1934, 1996.

Makalowski, W. and Boguski, M. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad Sci*, 95(16):9407–12, 1998.

Marcet-Houben, M. and Gabaldón, T. The Tree versus the Forest: The Fungal Tree of Life and the Topological Diversity within the Yeast Phylome. *PLoS One*, 4:8, 2009.

Marques, A. and Ponting, C. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biology*, 2009.

Matsen, F. A., Mossen, E., and Steel, M. A. Mixed-up trees: The structure of phylogenetic mixtures. *Bulletin of Mathematical Biology*, 70:1115–1139, 2008.

McLachlan, G. M. and Krishnan, T. *The EM Algorithm and Extensions.* Wiley Series in Probability and Statistics, 2008. ISBN-10: 0-471-20170-7.

Misof, B., Anderson, C. L., Buckley, T. R., Erpenbeck, D., Rickert, A., and Misof, K. An empirical analysis of mt 16S rRNA covarion-like evolution in insects: Site-specific rate variation is clustered and frequently detected. *J. Mol. Evol.*, 55(4):460–469, 2002.

Mossel, E. and Vigoda, E. Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science*, 309(5744):2207–2209, 2005.

Nagaki, Y., Susko, E., Fast, N., and Roger, A. Covarion shifts cause a long-branch attraction artefact that unites microsporidia and archaebacteria in ef-1α phylogenies. *Mol. Biol. Evol.*, 21(7):1340–1349, 2004.

Needleman, S. B. and Christian, W. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48 (3):443—53, 1970.

Nguyen, M. A. T., Klaere, S., and von Haeseler, A. MISFITS: Evaluating the goodness of fit between a phylogenetic model and an alignment. *Mol. Biol. and Evol.*, 28(1): 143–52, Jan 2011.

Nikolaev, S., Montoya-Burgos, J. I., Margullies, E. H., ComparativeSequencingProgram, Rougemont, J., Nyffeler, B., and Antonarakis, S. E. Early history of mammals is elucidated with the encode multiple species sequencing data. *PLoS Genetics*, 3, 1 e2, 2007.

Notredame, C., Higgins, D. G., and Heringa, J. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J.Mol.Biol.*, 302:205–217, 2000.

Orban, T. and Olah, E. Purifying selection on silent sites: a constraint from splicing regulation? *Trends Genet.*, 17:252–253, 2001.

Pachter, L. and Sturmfels, B., editors. *Algebraic Statistics for computational biology.* Cambridge University Press, November 2005a. ISBN 0-521-85700-7.

Pachter, L. and Sturmfels, B., editors. *Algebraic Statistics for Computational Biology.* Cambridge University Press, November 2005b. ISBN 0-521-85700-7.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40 (12):1413–5, 2008.

Parmley, J., Chamary, J., and Hurst, L. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. and Evol*, 23:301–309, 2006.

Pearson, W. R. and Lipman, D. J. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the USA*, 85:2444—-8, 1988.

Penny, D., McComish, B., Charleston, M., and Hendy, M. Mathematical elegance with biochemical realism: The covarion model of molecular evolution. *J. Mol. Evol.*, 53: 711–723, 2001.

Pheasant, M. and Mattick, J. Raising the estimate of functional human sequences, 2007.

Pistone, G. Generalised confounding with Grobner bases. *Biometrika*, 83(3):653–666, September 1996. ISSN 0006-3444.

Pistone, G., Riccomagno, E., and Wynn, H. *Algebraic Statistics: Computational Commutative Algebra in Statistics.* Chapman & Hall/CRC, December 2000.

Pollard, D. A., Iyer, V. N., Moses, A. M., and Eisen, M. B. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. *PLoS genetics*, 2(10):e173, 2006a.

Pollard, K., Salama, S., King, B., Kern, A., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J., Bejerano, G., Baertsch, R., and et al. Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics*, 2 (10), 2006b. e168.

Pollard, K., Hubisz, M., Rosenbloom, K., and Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

Ponjavic, J., Ponting, C., and Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.*, 17(5):556—-65, 2007.

Posada, D. jModelTest: phylogenetic model averaging. *Mol. Biol. and Evol.*, 25(7): 1253–1256, 2008.

Posada, D. and Crandall, K. A. Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Mol. Biol. and Evol.*, 18(6): 897–906, 2001.

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

Rao, C. *Linear statistical inference and its applications.* Wiley, NY, 1973.

Rhodes, J. and Sullivant, S. Identifiability of large phylogenetic mixture models. http://arxiv.org/abs/1011.4134v1, 2011.

Riccomagno, E. A short history of algebraic statistics. *Metrika*, 69(2-3):397–418, 2008. ISSN 00261335.

Ripplinger, J. and Sullivan, J. Does choice in model selection affect maximum likelihood analysis? *Systematic Biology*, 57(1):76–85, 2008.

Ripplinger, J. and Sullivan, J. Assessment of Substitution Model Adequacy Using Frequentist and Bayesian Methods. *Mol. Biol. and Evol.*, 27(12):2790–2803, 2010.

Ronquist, F., Larget, B., Huelsenbeck, J. P., Kadane, J. B., Simon, D., and van der Mark, P. Comment on "phylogenetic mcmc algorithms are misleading on mixtures of trees". *Science (N.Y.)*, 312(5772):367; author reply 367, Apr 2006.

Rotman, J. *An introduction to the theory of groups.* Springer, 1995.

Schwartz, R. S. and Mueller, R. L. Branch length estimation and divergence dating: estimates of error in Bayesian and maximum likelihood frameworks. *BMC Evolutionary Biology*, 10(1):5, 2010.

Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2): 461–464, March 1978.

Semple, C. and Steel, M. *Phylogenetics*, volume 24 of Oxford Lecture Series in Mathematics and its Applications. Oxford University Press, Oxford, 2003. ISBN 0-19-850942-1.

Serre, J. *Linear representations of finite groups.* Springer-Verlag, New York, 1977. ISBN 0-387-90190-6. Translated from the second French edition by Leonard L. Scott, Graduate Texts in Mathematics, Vol. 42.

Shabalina, S. and Kondrashov, A. Pattern of selective constraint in C. elegans and C. briggsae genomes. *Genet Res.*, 74(1):23–30, 1999.

Shabalina, S., Ogurtsov, A., Lipman, D., and Kondrashov, A. Patterns in interspecies similarity correlate with nucleotide composition in mammalian 3'UTRs. *Nucleic Acids Res.*, 31(18):5433—-5439, 2003.

Siepel, A., Bejerano, G., Pedersen, J., Hinrisch, A., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L., Richards, S., and et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034—50, 2005.

Spicher, A., Guicherit, O., Duret, L., Aslanian, A., Sanjines, E., Denko, N., Giaccia, A., and Blau, H. Highly conserved RNA sequences that are sensors of environmental stress. *Mol. Cell Biol.*, 18(12):7371–82, 1998.

Steel, M. A., Hendy, M. D., Székely, L. A., and Erdös, P. L. Spectral analysis and a closest tree method for genetic sequences., 1992. ISSN 0893-9659.

Steel, M. A., Székely, L. A., and Hendy, M. D. Reconstructing trees when sequence sites evolve at variable rates. *Journal of Computational Biology*, 1(2):153–163, 1994.

Steel, M. A., Huson, D., and Lockhart, P. J. Invariable sites models and their use in phylogeny reconstruction. *Systematic Biology*, 49:225–232, 2000.

Stefankovic, D. and Vigoda, E. Pitfalls of heterogeneous processes for phylogenetic reconstruction. *Systematic biology*, 56(1):113–24, February 2007.

Stefanovic, D. and Vigoda, E. Phylogeny of mixture models: Robustness of maximum likelihood and non-identifiable distributions. *J. Comput. Biol.*, 14:156–189, 2007.

Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct & Mol. Biol*, 14(2):103—-5, 2007.

Sturmfels, B. and Sullivant, S. Toric ideals of phylogenetic invariants. *J. Comput. Biol.*, 12:204–228, 2005.

Sugnet, C. W., Srinivasan, K., Clark, T. A., O'Brien, G., Cline, M. S., Wang, H., Williams, A., Kulp, D., Blume, J. E., Haussler, D., and Ares, J. Unusual Intron Conservation near Tissue-Regulated Exons Found by Splicing Microarrays. *PLoS Comput Biol*, 2(1):e4, 01 2006.

Sullivan, J. and Swofford, D. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Syst. Biol.*, 50(5):723–729, 2001.

Sullivant, S. Algebraic geometry of Gaussian Bayesian networks. *Adv. in Appl. Math.*, 40(4), 2008.

Swofford, D. L. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). *Sinauer Associates*, 2003. Version 4.

Swofford, D., Waddell, P., Huelsenbeck, J., Foster, P., Lewis, P., and Rogers, J. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.*, 50(4):525–539, 2001.

Tanner, M. A. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, volume 3rd Edition. Springer Series in Statistics, New York: Springer, 1996.

Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. In *Some mathematical questions in biology—DNA sequence analysis (New York, 1984)*, volume 17 of *Lectures Math. Life Sci.*, pages 57–86. Amer. Math. Soc., Providence, RI, 1986.

TheENCODEProjectConsortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, 2007.

TheENCODEProjectConsortium. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol*, 9(4), 2011. e1001046.

Thomas, D. J., Rosenbloom, K. R., Clawson, H., Hinrichs, A. S., Trumbower, H., Raney, B. J., Karolchik, D., Barber, G. P., Harte, R. A., Hillman-Jackson, J., and et.al. The ENCODE Project. *Nucleic Acids Research*, 35:D663-7, Jan 2007.

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

Tilgner, H. and Guigó, R. Improving splicing prediction accuracy helps to find putative splicing regulators. Eukaryotic mRNA processing, 2007.

Tilgner, H. and Guigó, R. From chromatin to splicing: Rna-processing as a total artwork. *Epigenetics : official journal of the DNA Methylation Society*, 5(3)(3):180–184, Apr 2010.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., and Guigó, R. Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.*, 16(9):996–1001, Sep 2009.

Tuffley, C. and Steel, M. Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.*, 147(1):63–91, 1998.

Venter, J. C., Adams, M D ad Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., and et.al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

Warnecke, T. and Hurst, L. D. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster. *Mol. Biol. and Evol.*, 24(12):2755–2762, 2007.

Warnecke, T., Batada, N., and Hurst, L. D. The Impact of the Nucleosome Code on Protein-Coding Sequence Evolution in Yeast. *PLoS Genetics*, 4(11):12, 2008a.

Warnecke, T., Parmley, J., and Hurst, L. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biology*, 9(2):R29, 2008b.

Warnecke, T., Weber, C., and Hurst, L. Why there is more to protein evolution than protein function: splicing, nucleosomes and dual-coding sequence. *Biochemical Society Transactions*, 37(4):756–761, 2009.

Washietl, S., Machné, R., and Goldman, N. Evolutionary footprints of nucleosome positions in yeast. *Trends in genetics : TIG*, 24(12):583–7, December 2008. ISSN 0168-9525. doi: 10.1016/j.tig.2008.09.003.

Wegrzyn, J., Drudge, T., Valafar, F., and Hook, V. Bioinformatic analyses of mammalian 5'-UTR sequence properties of mRNAs predicts alternative translation initiation sites. *BMC Bioinformatics*, 9(1):232, 2008.

Whelan, S. and Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, pages 18691—-699, 2001.

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

Wu, C. On the Convergence Properties of the EM Algorithm. *Ann. Statist.*, 11(1): 95–103, 1983.

Xiao, X., Wang, Z., Jang, M., and Burge, C. B. Coevolutionary networks of splicing cis-regulatory elements. *PNAS*, 104(47), November 2007.

Xing, Y. and Lee, C. Alternative splicing and RNA selection pressure—Evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.*, 7:499—-509, 2006.

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

Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. and Evol.*, 24(8):1586–1591, 2007. ISSN 07374038.

Yap, V. B. and Pachter, L. Identification of evolutionary hotspots in the rodent genomes. *Genome research*, 14(4):574–9, April 2004. ISSN 1088-9051.

Zacks, S. *The theory of statistical inference.* John Wiley and Sons, Inc., New York, 1971.

Zhang, Z. and Gerstein, M. Large-scale analysis of pseudogenes in the human genome. *Curr. Opin. Genet. Dev.*, 14:328—-335, 2004.

Zou, L., Susko, E., Field, C., and Roger, A. The Parameters of the Barry and Hartigan General Markov Model Are Statistically NonIdentifiable. *Syst. Bio*, 60:872–875, 2011.