Probabilistic Models for Molecular Phylogenetics and Systems Biology

Dirk Husmeier

Habilitationsschrift

A thesis submitted
in conformity with the requirements
for the degree of a higher doctorate
(Habilitation)

Department of Statistics
TU Dortmund University
44221 Dortmund
Germany

Edinburgh, March 2011

Acknowledgements

Writing this thesis would not have been possible without the diligent work of my (former) students Adriano Werhli, Wolfgang Lehrach, Anna Kedzierska, Frank Dondelinger, Ali Faisal and Alex Mantzaris, as well as my former postdocs Marco Grzegorczyk and Kuang Lin. I am indebted to my senior colleague Frank Wright, who introduced me to bioinformatics and phylogenetics when I was still a novice in this field. I would also like to thank my line manager, Chris Glasbey, as well as my other colleagues at BioSS, who have not only contributed to turning this institute into a stimulating research environment, but have also encouraged me to take a more prudent approach to my work-life balance. The work described in the first three chapters was funded by the Biotechnology and Biological Sciences Research Council (BBSRC). The remaining work was funded by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD).

Preface

This thesis provides an overview of the research I have carried out at Biomathematics & Statistics Scotland (BioSS) during the last eleven years. I have divided it into two parts, covering the two main areas of my contributions: statistical phylogenetics and computational systems biology. The objective is to present a high-level overview of my research, so as to outline its main ideas, describe the underlying methodological concepts, put the work into the context of other state of the art paradigms, and summarize its main findings. For the detailed mathematical derivations, algorithmic implementations, and simulation procedures, I refer the reader to my original publications. In most chapters I have included particular examples of data analysis. These are to be taken as illustrative rather than exhaustive, to give the reader a feel for typical application areas. Comprehensive comparative evaluations are mostly available from my papers, even if not specifically pointed out in the text. I have largely refrained from including detailed mathematical descriptions. The equations that I have included are mostly of a general nature, so as to sketch the essential concepts and methodological framework. Again, the reader is referred to my original papers for the specific details. These papers are cited by number, with numbers included in square brackets. This is to be distinguished from equations, which are labelled by numbers in standard brackets. To distinguish my own articles from citations of other researchers' papers, I precede the former by my initials. As an example, [DH12] refers to the twelfth article in my publication list, while [12] refers to the twelfth article in the general bibliography. Both lists are included as separate sections at the end. For environmental reasons, I have not attached my papers as hard copies to this document. Instead, I have made them available as PDF files on the following website: http://www.bioss.ac.uk/∼dirk/SelectedPublications. The names of the files are identical to the citation labels, except for my book [DH1], which I have divided into chapters. For instance, to obtain the twelfth paper from my publication list, one has to download file DH12.pdf. The second chapter of my book is available in file DH1_chapter2.pdf.

My hope is that this thesis is a useful document that can stand on its own, to review the 'wood' of general themes and concepts without climbing the 'trees' of methodological and implementational details. The prospective target audience are statisticians and machine learning researchers with an interest in postgenomic biology, as well as biologists with an interest in quantitative methods. However, to benefit form the first part of this thesis, a certain background in phylogenetics is required. This does not only include a familiarity with the concept of a phylogenetic tree, but also requires an understanding of the three established paradigms of phylogenetic inference: clustering based on pairwise sequence distances, maximum parsimony, and likelihood methods. To keep this document self-contained, I have therefore included, in Chapter 1, a brief review of statistical phylogenetics. The remaining chapters describe my own research.

Synopsis

The final decade of the last millennium witnessed a dramatic revolution in molecular biology. This revolution was heralded by a large-scale DNA sequencing effort in 1995, when the entire 1.8 million base pairs of the genome of the bacterium *Haemophilus influenzae* was published. Since then, the amount of DNA sequences in publicly accessible data bases has been growing exponentially, including the complete 3.3 billion base-pair DNA sequence of our own species. In addition, new high-throughput technologies like microarrays and yeast 2-hybrid assays have resulted in a deluge of further data, related to the transcriptome, the spliceosome, the proteome, and the metabolome. This "omic" explosion has led to a paradigm shift in molecular biology. While the pre-genomic era was dominated by hypothesis-driven reductionist approaches, founded on qualitative methods and characterized by a certain ingrained scepticism about the merits of mathematical modelling, post-genomic molecular biology takes a holistic, systems-based approach, which is data-driven and increasingly relies on quantitative methods. The problems faced are essentially statistical, due to the inherent complexity and stochasticity of biological systems, the random processes intrinsic to evolution, and the unavoidable error-proneness and variability of high-throughput postgenomic assays. In the present thesis, I shall review the contributions I have made to this field during the last eleven years, ever since starting my employment as a researcher in "statistical bioinformatics" at BioSS.

Part 1 will focus on pattern recognition in DNA sequence alignments within a phylogenetic context. I will start off, in **Chapter 1**, with a brief review of molecular phylogenetics, based on [DH1], and then discuss the problem of interspecific recombination. The underlying assumption of most phylogenetic tree reconstruction methods is that there is one set of hierarchical relationships among the taxa. While this is a reasonable approach when applied to most DNA sequence alignments, it can be violated in certain bacteria and viruses due to interspecific recombination. The resulting transfer or exchange of DNA subsequences can lead to a change of the branching order (topology) in the affected region, which results in conflicting phylogenetic information from different regions of the

alignment. If undetected, the presence of these so-called mosaic sequences can lead to systematic errors in phylogenetic tree estimation. Their detection, therefore, is a crucial prerequisite for consistently inferring the evolutionary history of a set of DNA sequences.

Summarizing my work in [DH11,DH17,DH21,DH22,DH29], **Chapter 2** describes a graphical exploratory method based on moving a window along the DNA sequence alignment and computing a divergence measure of a phylogenetic signal as a potential indicatior of evidence for recombination. The phylogenetic signal chosen is the posterior distribution of phylogenetic tree topologies, which is based on the Bayesian approach to phylogenetic inference, and which is practically computed by repeatedly running Markov chain Monte Carlo (MCMC) simulations on different subsets of the DNA sequence alignment [DH29]. This method can be refined with a topology based pruning scheme [DH21] and the combination with a Bayesian hidden Markov model [DH17].

Chapter 3 (summarizing my work in [DH15,DH25,DH26,DH28]) describes an approach based on phylogenetic HMMs. The method is based on the observation that interspecific recombination usually leads to a change of the underlying phylogenetic tree topology. The idea is to introduce a hidden state that represents the tree topology at a given site. A state transition from one topology into another corresponds to a recombination event. To introduce correlations between adjacent sites, the hidden states are given a Markovian dependence structure. Thus, the standard model of a phylogenetic tree is generalized by the combination of two probabilistic models: a taxon graph (phylogenetic tree) representing the relationships among the taxa, and a site graph (HMM) representing dependencies between different sites in the DNA sequence alignments. Changepoints of mosaic segments in the alignment are predicted by state transitions in the site graph. While this method can only deal with a small number of sequences simultaneously, it has the potential to predict the locations and changepoints of recombinant regions more accurately than what can be achieved with most existing techniques.

A shortcoming of phylogenetic HMMs is their inability to differentiate between recombination and rate heterogeneity. I address this issue in **Chapter 4** (summarizing my work in [DH9,DH10,DH20,DH46]) by extending the HMM approach to a factorial HMM (FHMM). The states of the first hidden chain represent tree topologies, as before, and transitions between these states are indicative of recombination events. The states of the second independent hidden chain represent different global scaling factors of the branch lengths, and transitions between these rate states indicate rate heterogeneity, potentially related to variations in the selective pressure. I discuss different Bayesian inference schemes based on Gibbs sampling [DH20] and transdimensional reversible jump MCMC [DH9], Bayesian model selection [DH10] and within-codon effects [DH46].

Part 2 will be devoted to systems biology. I start, in **Chapter 5** (summarizing my work in [DH19,DH50]) with protein-protein interactions involved in the formation of macromolecular complexes and biochemical pathways. Since high-throughput experiments like yeast two-hybrid and phage display are expensive and intrinsically noisy, it would be desirable to more specifically tar-

get or partially bypass them with complementary in silico approaches. The chapter presents a probabilistic discriminative approach to ab initio prediction of protein-peptide interactions from sequence data and discusses how susceptibility to overfitting can be avoided by adopting a Bayesian a posteriori approach based on a Laplacian prior in parameter space.

Molecular pathways consisting of interacting proteins underlie the major functions of living cells, and a central goal of systems biology is the elucidation of their structure and regulatory mechanisms. Summarizing my work in [DH3,DH4,DH5,DH12,DH13,DH14,DH16,DH18,DH23,DH41,DH42,DH44,DH45], **Chapter 6** covers the reconstruction of gene regulatory networks from transcriptomic profiles with probabilistic graphical models. It discusses the evaluation of the network reconstruction accuracy from a realistic simulation study [DH18,DH23], the systematic integration of prior knowledge in a Bayesian framework [DH13,DH16], the inference of evolving network structures during an organism's life cycle [DH41,DH42], the approximate modelling of nonlinear regulation [DH12,DH44,DH45], and improved MCMC schemes [DH3,DH4,DH14] for Bayesian learning of Bayesian networks.

Transcriptional gene regulation is primarily controlled by diverse regulatory proteins called transcription factors (TFs), which bind to specific DNA sequences and thereby repress or initiate gene expression. The regulation thus depends on (TF) activities, that is the concentration of the TF subpopulation capable of DNA binding. The methods described in Chapter 6 approximate the activities of TFs by their gene expression levels. However, TFs are frequently subject to post-transcriptional and post-translational modifications, which may affect their DNA binding capability. Consequently, gene expression levels of TFs may only contain limited information about their actual activities. In **Chapter 7** (summarizing my work in [DH8]) I discuss a method that addresses this difficulty by treating TFs as latent variables and inferring their activity profiles along with the regulatory network structure from microarray and immunoprecipitation data by application of the variational Bayesian expectation maximization algorithm.

The thesis concludes in **Chapter 8** with an outlook on how methods from computational systems biology can be applied and adapted to modern ecology. Summarizing my work in [DH5], the chapter discusses the methodological hurdles that have to be overcome to infer species interaction networks from species distribution data. Understanding these networks is of growing importance for the protection of biodiversity, as they help us to predict potential knock-on effects of a biological control agent, or to understand how species will respond to climate change.

Part I

Molecular Phylogenetics

Chapter 1

Introduction

This chapter provides a brief review of molecular phylogenetics, based on [DH1]. I introduce the concepts of rooted and unrooted phylogenetic trees and summarize the key ideas and relative merits of the three established paradigms of phylogenetic inference: clustering based on pairwise sequence distances, maximum parsimony, and likelihood methods. I will then discuss the problem of interspecific recombination. The underlying assumption of most phylogenetic tree reconstruction methods is that there is one set of hierarchical relationships among the taxa. While this is a reasonable approach when applied to most DNA sequence alignments, it can be violated in certain bacteria and viruses due to interspecific recombination. The resulting transfer or exchange of DNA subsequences can lead to a change of the branching order (topology) in the affected region, which results in conflicting phylogenetic information from different regions of the alignment. If undetected, the presence of these so-called mosaic sequences can lead to systematic errors in phylogenetic tree estimation. Their detection, therefore, is a crucial prerequisite for consistently inferring the evolutionary history of a set of DNA sequences.

1.1 Background on Phylogenetic Trees

The objective of phylogenetics is to reconstruct the evolutionary relationships among different species or strains (generic name *taxa*[1]) and to display them in a binary or bifurcating tree-structured

[1]In bacteria and viruses it is difficult to distinguish between *species* and *strains*. This chapter is rather sloppy in the use of these terms, and occasionally uses both terms synonymously.

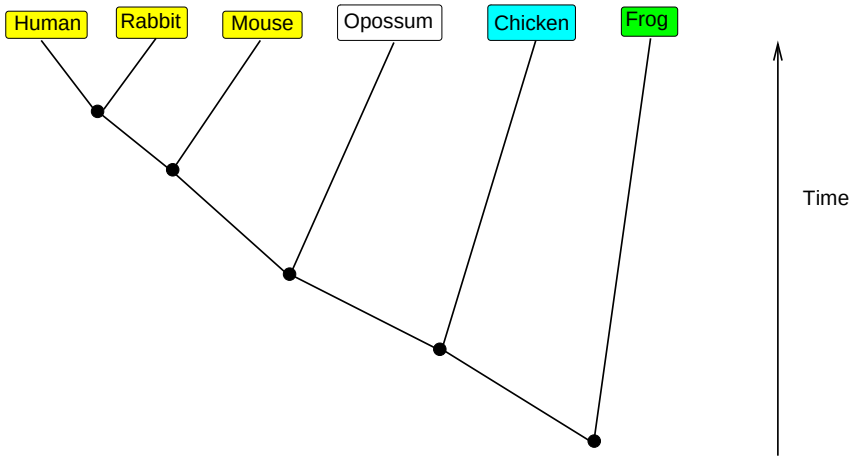


Figure 1.1: **Rooted phylogenetic tree.** Leaf nodes represent extant or contemporary species. Hidden nodes represent hypothetical ancestors, where lineages bifurcate (so-called *speciation* events).

graphical model called a *phylogenetic tree*. An example is given in Figure 1.1. The leaves of a phylogenetic tree represent contemporary species, like chicken, frog, mouse, etc. The inner or hidden nodes represent hypothetical ancestors, where a splitting of lineages occurs. These so-called speciation events lead to a diversification in the course of evolution, separating, for instance, warm-blooded from cold-blooded animals, birds from mammals, primates from rodents, and so on.

Figure 1.1 shows a so-called *rooted tree*. The node at the bottom of the tree represents the most recent common ancestor, from which ultimately all other species descended. The edges are directed, related to the direction of time: the closer a node is to the root of the tree, the older is the corresponding speciation event. Inferring this direction of evolutionary processes, however, is difficult and not amenable to most modelling and inference methods. Consequently, phylogenetic trees will be displayed as *unrooted trees*, shown in Figure 1.2. As opposed to a rooted tree, the edges in an unrooted tree are undirected, and no node is in the distinguished position of a root.[2]

A phylogenetic tree conveys two types of information. The *topology* defines the branching order of the tree and the way the contemporary species are distributed among the leaves. For example, from Figures 1.1 and 1.2 we learn that the mammals – human, chicken, mouse, and opossum – are grouped together and are separated from the group of animals that lay eggs – chicken and frog. Within the former group, opossum is grouped out because it is a marsupial and therefore

[2]We can regain a rooted from an unrooted tree by including an *outgroup* in the analysis. An outgroup is a (set of) species that is known *a priori* to be less related to any of the other taxa used in the study, and the root will therefore be located on the branch between the other taxa and the outgroup. For example, in Figure 1.2, *frog* is the outgroup, because it is the only cold-blooded animal among a set of warm-blooded animals. By positioning the root on the branch leading to frog, we regain the rooted tree of Figure 1.1 (although the exact position of the root on this branch is not known).

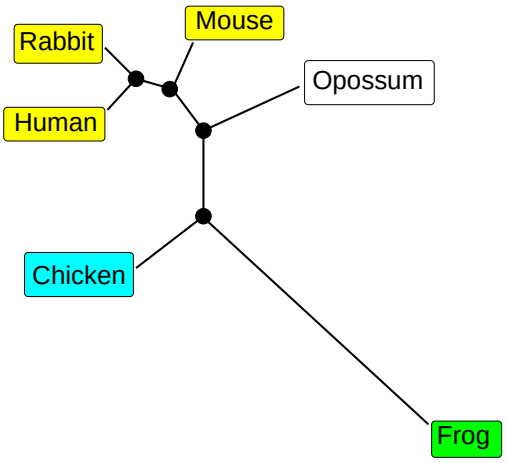


Figure 1.2: **Unrooted phylogenetic tree.** Leaf nodes represent contemporary species, hidden nodes represent hypothetical ancestors. The tree conveys two types of information. The *topology* defines the branching order of the tree and the way the contemporary species are distributed among the leaves. The *branch lengths* represent phylogenetic time, measured by the average amount of mutational change.

less closely related to the other "proper" mammals. Exchanging, for instance, the leaf positions of opossum and rabbit changes the branching order and thus leads to a different tree topology. For n species there are, in total, $(2n-3)!!$ different rooted, and $(2n-5)!!$ different unrooted tree topologies, where !! denotes double factorial: $(2n-5)!! = (2n-5)(2n-7)...1$. For a proof, see e.g. Chapter 4 in [DH1]. In what follows, we will use the integer variable $S \in \{1, 2, \ldots, (2n-5)!!\}$ to label the different unrooted tree topologies.

The second type of information we obtain from a phylogenetic tree is given by the *branch lengths*, which represent *phylogenetic time*,[3] measured by the average amount of mutational change. For example, Figure 1.2 shows a comparatively long branch leading to the leaf with *frog*. This long branch indicates that a comparatively large number of mutations separates *frog* from the other animals, and that a large amount of phylogenetic time has passed since the lineage leading to *frog* separated from the remaining tree. This conjecture is reasonable because *frog* is the only cold-blooded animal, whereas all the other animals are warm-blooded. Note that the mutation rate is usually unknown, and we can therefore not infer *physical time* from the branch lengths. I will revisit this distinction between *physical* time and *phylogenetic* time in Section 1.4; see equation (1.1).

An unrooted tree for n species has $n-2$ inner nodes, and thus $m = n + (n-2) - 1 = 2n-3$ branches. In what follows, individual branch lengths will be denoted by w_i, and the total vector of branch lengths will be denoted by $\mathbf{w} = (w_1, \ldots, w_{2n-3})$.

[3] I use *phylogenetic time* with the meaning *evolutionary change*, which is given by the product of *physical time* and a *nucleotide substitution rate*. See equation (1.1) for an exact definition.

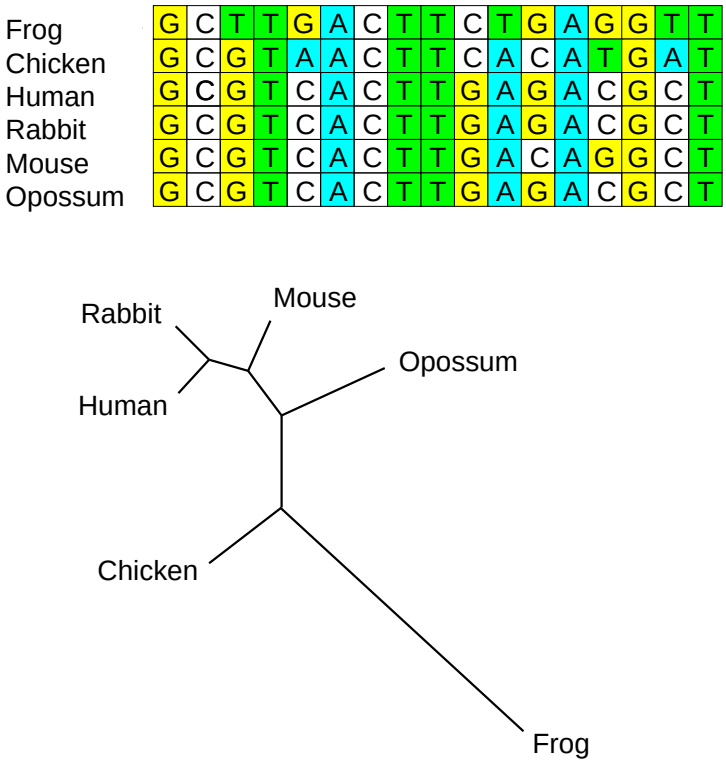


Figure 1.3: **Phylogenetic inference from DNA sequence alignments.** The *top* figure shows a section of the DNA sequence alignment for the β-globin encoding gene in six species. The total alignment has a length of 444 nucleotides. The *bottom* figure shows a phylogenetic tree inferred from this alignment with the method of maximum likelihood, to be discussed in Section 1.4.

The objective of statistical phylogenetics is to develop methods for reconstructing phylogenetic trees. Since the driving forces of evolution are *mutations*, that is, errors in the *replication* of DNA, it seems reasonable to base the inference on a comparison of DNA sequences[4] obtained from the different species or strains of interest. This approach has become viable by major breakthroughs in DNA sequencing techniques, with the number of DNA sequences in publicly accessible data bases growing exponentially.

DNA is composed of an alphabet of four *nucleotides*, which come in two families: the purines *adenine* (A) and *guanine* (G), and the pyrimidines *cytosine* (C) and *thymine* (T). DNA sequencing is the process of determining the order of these nucleotides. After obtaining the DNA sequences of the taxa of interest, we need to compare *homologous* subsequences, that is, corresponding regions of the genome that code for the same protein. More precisely, we have to compare homologous

[4]Earlier approaches to phylogenetics were based on the analysis of non-molecular data, for instance, morphological characters. This chapter focuses on phylogenetic analysis based on molecular sequence data, mainly DNA sequences. The advantages of this more recent approach of *molecular phylogenetics* over classical *non-molecular phylogenetics* are discussed in [103].

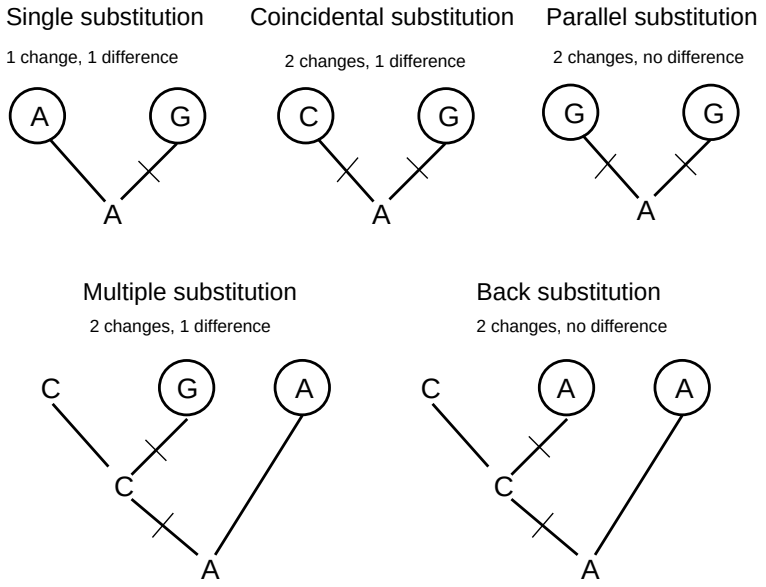


Figure 1.4: **Hidden and multiple nucleotide substitutions.** The figure contains several phylogenetic trees, which display different nucleotide substitution scenarios. In each tree, the circles at the leaf level represent contemporary sequences from a DNA sequence alignment, for which a pairwise distance is to be computed. The letters inside these circles represent nucleotides at a given site in the alignment. Nodes further down in the tree hierarchy (that is, below the leaf nodes) represent (usually extinct) ancestors. The letters at these nodes show nucleotides at the corresponding site in the ancestral DNA sequence. For the tree in the top left, the observed number of nucleotide substitutions is identical to the actual number of substitutions that have occurred during evolution. However, in all the other cases, the actual number of nucleotide substitutions is larger than the observed number, hence a naive approach to computing pairwise evolutionary distances systematically underestimates their true values.

nucleotides, that is, nucleotides which have been acquired directly from the common ancestor of the taxa of interest. The process of establishing which regions of a set of DNA sequences are homologous and should be compared is called *DNA sequence alignment*. Throughout this thesis I will assume that the sequence alignment is given. For more comprehensive review, see [36].

The top of Figure 1.3 shows a small section of the DNA sequence alignment obtained from the β-globin encoding genes in six species (four mammals, one bird, and one amphibian). Rows represent different species, columns represent different sites or positions in the DNA sequence alignment. At the majority of sites, all nucleotides are identical, reflecting the fact that the sequences compared are homologous. At certain positions, however, differences occur, resulting from mutational changes during evolution. In the fifth column, for instance, human, rabbit, mouse, and opossum have a C, chicken has an A, and frog has a G. This reflects the fact that the first four species are mammals and therefore more closely related to each other than to the two remaining species. Obviously, however, nucleotide substitutions are not as deterministically related to the phylogenetic tree as in this example. Evolution is driven by stochastic forces that act on genomes, and our objective is to

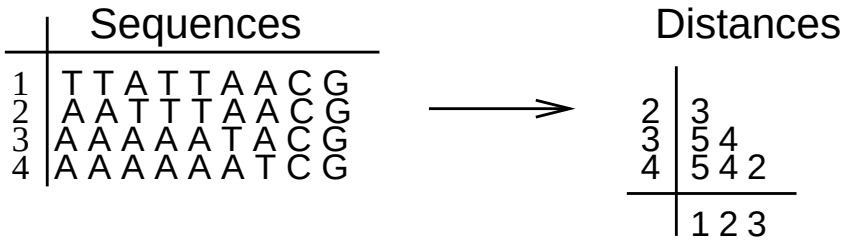


Figure 1.5: **Information loss of distance methods.** By computing pairwise evolutionary distances from a DNA sequence alignment, the high-dimensional space of DNA sequences is mapped into the low-dimensional space of pairwise distances. This substantial dimension reduction leads to an inevitable loss of information.

discern significant similarities between diverged sequences amidst a chaos of random mutation and natural selection. Faced with noisy data resulting from intrinsically stochastic processes, the most powerful methods make use of probability theory. Based on a mathematical model of nucleotide substitution in terms of a homogeneous Markov chain, a phylogenetic tree can be interpreted as a probabilistic generative model, which allows us to compute the likelihood of an observed DNA sequence alignment. The practical computation draws on well-established inference algorithms for Bayesian networks[5], which allows the application of standard inference procedures from statistics and machine learning[6]. In Section 1.4 I provide a brief review of probabilistic model-based inference in phylogenetics. I shall motivate this approach by first briefly describing the shortcomings of the older approaches of *sequence distance-based clustering* and *maximum parsimony*.

1.2 Clustering based on Pairwise Sequence Distances

A straightforward way to get a phylogenetic tree from a DNA sequence alignment is as follows: (1) count the number of nucleotides by which two DNA sequences differ; (2) repeat this procedure for all pairs of sequences and compute a table of pairwise distance scores; (3) apply a clustering algorithm, like Neighbour Joining [124, 138], to obtain a dendrogram. The approach appears attractive because it avoids the NP-hardness of parsimony and likelihood methods, discussed below. However, it suffers from three fundamental shortcomings. First, there is no guarantee that the resulting tree is optimal in any sense, as the method does not aim to optimize any objective function. Second, by counting the number of different nucleotides, the true number of nucleotide substitution events is systematically underestimated and requires a (heuristic) correction; see Figure 1.4 for an illustration. Third, the method effectively maps the high-dimensional space of DNA sequences into the low-dimensional space of pairwise distances, as illustrated in Figure 1.5, and this dimension reduction incurs a substantial loss of information. While being computationally cheap, clustering methods based on pairwise sequence distances are thus intrinsically suboptimal and should be

[5]See e.g. Chapter 2 in [DH1] for a review.
[6]See e.g. Chapter 1 in [DH1] for a review.

avoided if at all possible.

1.3 Maximum Parsimony

The basic idea behind *maximum parsimony*, introduced in [44], is that the optimal phylogenetic tree is the one requiring the smallest number of nucleotide substitutions along its branches. Take, as an example, Figure 1.6, which shows an alignment of four DNA sequences and a hypothetical phylogenetic tree, in which strain 1 is grouped with strain 2, and strain 3 is grouped with strain 4. We pick a column of the alignment and distribute the nucleotides across the leaf nodes. Next, we try to reconstruct the evolutionary history by assigning nucleotides to the hidden nodes, which correspond to ancestral, unobserved sequences. For the first column of the alignment, strains 1 and 2 have an A, while strains 3 and 4 have a C. If we assign nucleotides to the hidden nodes arbitrarily, say a G to the first hidden node, and a T to the second, as shown on the right of Figure 1.6, we can get up to five nucleotide substitution events needed for a reconstruction of evolution. However, a more skilful choice, shown at the bottom of Figure 1.6, can reduce this number to a single nucleotide substitution. Under the principle of parsimony this reconstruction is preferred over the less parsimonious reconstruction that requires five substitutions.

Next, we have to compare different trees. With four taxa, we have three unrooted tree topologies: strain 1 can either be grouped with strain 2, with strain 3, or with strain 4. For a given column of nucleotides from the alignment, we repeat the previous process of assigning nucleotides to the hidden nodes in the most parsimonious way. This is repeated for each of the possible tree topologies in turn. We then count the number of nucleotide substitutions needed. An example is given in Figure 1.7 where, for the first column in the alignment, the tree that groups strains 1 and 2 against strains 3 and 4 requires the smallest number of substitutions and is therefore preferred by parsimony. In fact, the whole process is repeated for all sites in the alignment. The total number of nucleotide substitutions is determined for each tree, and the tree that invokes the least total number of nucleotide substitutions is selected as the best candidate for the true evolutionary tree. Parsimony is not limited in the number of taxa, and the restriction to four sequences in the previous example was chosen for illustration purposes only. Note, however, that the number of tree topologies grows super-exponentially with the number of taxa and that the optimization problem is NP–hard [134]. Consequently, iterative and approximate methods are needed for large sequence alignments of many taxa.

As opposed to distance-based clustering, reviewed in the previous section, parsimony is an optimality method that minimizes a well-defined objective function: the total number of nucleotide substitutions along the tree branches. This approach shows a certain resemblance to minimum description length [113] in information theory and allows, unlike clustering methods, an evaluation

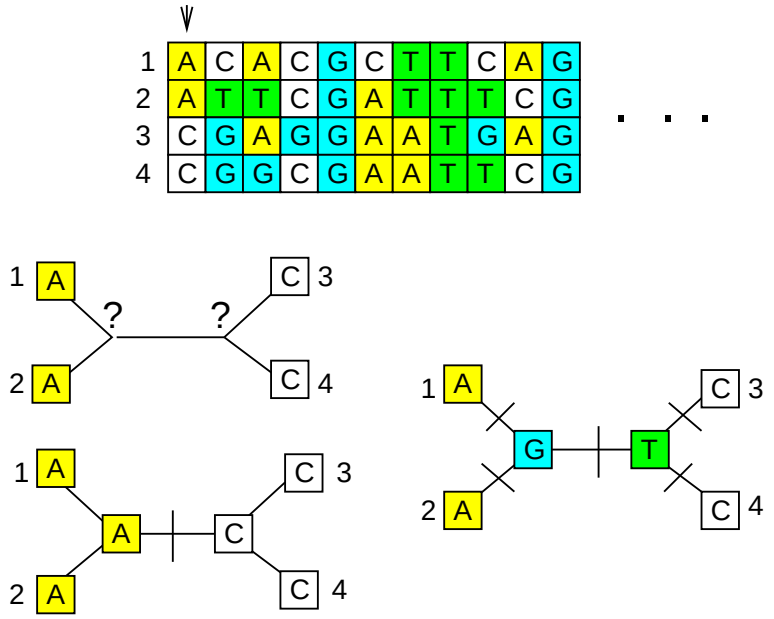


Figure 1.6: **Illustration of parsimony: single tree.** For each column of nucleotides in the alignment, the evolutionary history is to be reconstructed by assigning nucleotides to the hidden nodes. An arbitrary assignment, shown on the right, can give up to five nucleotide substitutions needed for a reconstruction of evolution. A more skilful choice, shown at the bottom, reduces this number to a single substitution. Under the principle of parsimony this reconstruction is preferred over the less parsimonious reconstruction that requires five substitutions.

of the quality of a reconstructed tree. However, as opposed to the methods to be described in the next section, parsimony is not based on a probabilistic generative model. Proponents of parsimony used to consider this model-free inference an advantage [43], but it has now become clear that it is, in fact, the limiting case of a model-based approach – for the limit of a highly implausible evolutionary scenario. For details, see Chapter 4 in [DH1].

The fundamental objection to parsimony is that it is not *consistent*. *Consistency* is the desirable feature of a method to converge to the right answer given enough data. In certain evolutionary scenarios, however, parsimony gives the wrong tree even if we add more and more data [40, 41]. The classic scenario where this might happen has been termed *long branch attraction* and is illustrated in Figures 1.8 and 1.9.

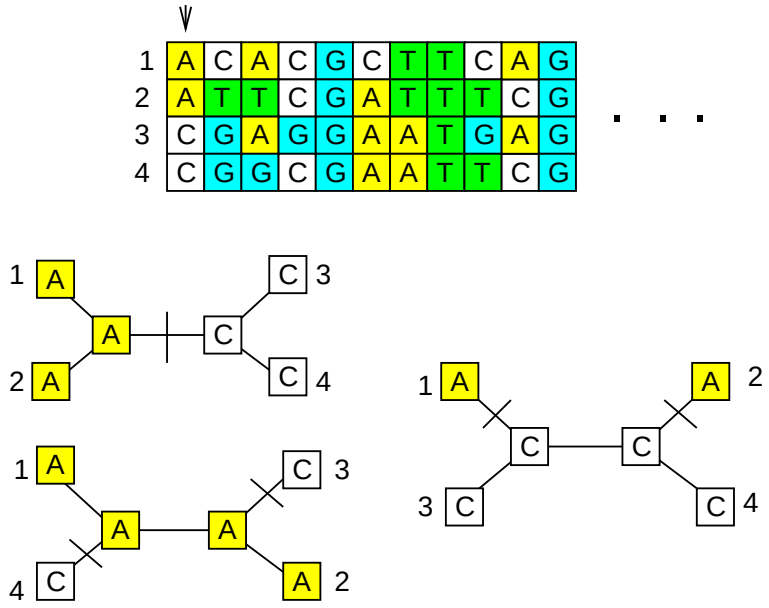


Figure 1.7: **Illustration of parsimony: tree selection.** The figure shows the three unrooted tree topologies that can be obtained with four taxa (where the taxa are represented by numbers). For a given site in the DNA sequence alignment, and for each of the possible tree topologies in turn, nucleotides are assigned to the hidden nodes in the most parsimonious way; compare with Figure 1.6. The tree in the top left, which invokes only a single substitution, is preferred over the other, less parsimonious trees, which require at least two substitutions. The whole process is repeated for all sites in the alignment, and the total number of nucleotide substitutions is counted, for each tree in turn. Then, the tree that invokes the least total number of nucleotide substitutions is selected as the best candidate for the true evolutionary tree.

1.4 Likelihood Methods

Likelihood methods are based on an explicit mathematical model of nucleotide substitution, which allows the formulation of the inference process in terms of a probabilistic generative model. This fact renders likelihood methods considerably more powerful than clustering based on pairwise sequence distances (Section 1.2) or the method of maximum parsimony (Section 1.3). As opposed to clustering, inference is based on an optimality function, which allows us to objectively compare different trees and to test hypotheses about evolutionary scenarios. Inference is based on the whole DNA sequence alignment, which avoids the information loss inherent in pairwise sequence distance methods (as illustrated in Figure 1.5). By basing the inference on an explicit mathematical model of nucleotide substitution, the shortcomings of "model-free" inference and the inconsistency of maximum parsimony are avoided. In particular, likelihood methods subsume maximum parsimony as a limiting case for a rather unrealistic evolutionary scenario[7].

[7]See e.g. Chapter 4 of [DH1] for details.

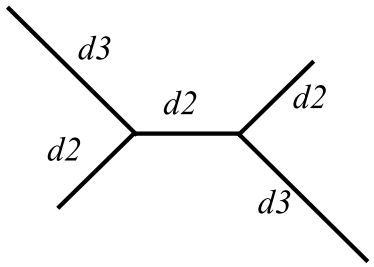


Figure 1.8: **Felsenstein zone.** The figure shows a phylogenetic tree of four taxa and two types of branch lengths, denoted by $d2$ and $d3$. For configurations with large branch lengths $d3$ and small branch lengths $d2$ (the so-called 'Felsenstein zone'), the method of maximum parsimony is known to systematically infer the wrong tree topology, as shown in [40]. An explanation is given in Figure 1.9.

The driving forces for evolution are nucleotide substitutions, which can be modelled as transitions in a 4-element state space, shown in Figure 1.10. $P(y|x,w)$, where $x, y \in \{A, C, G, T\}$, denotes the probability of a transition from nucleotide x into nucleotide y, conditional on the elapsed phylogenetic time w. The latter is given by the product of an unknown nucleotide substitution rate α with physical time t:

$$w \;=\; \alpha t \tag{1.1}$$

To rephrase this: $P(y|x,w)$ is the probability that nucleotide y is found at a given site in the DNA sequence given that w phylogenetic time units before the same site was occupied by nucleotide x.

An intuitively plausible functional form for these probabilities is shown on the right of Figure 1.10. For $w = 0$, there is no time for nucleotide substitutions to occur. Consequently, $P(A|A, w = 0) = 1$, and $P(C|A, w = 0) = P(G|A, w = 0) = P(T|A, w = 0) = 0$. As w increases, nucleotide substitutions from A into the other states lead to an exponential decay of $P(A|A, w)$ and, concurrently, an increase of $P(C|A, w)$, $P(G|A, w)$, and $P(T|A, w)$. The rate of this decay or increase depends on the type of nucleotide substitution, as illustrated in Figure 1.11. Nucleotides are grouped into two families: purines (A and G), and pyrimidines (C and T). Nucleotide substitutions within a nucleotide class (purine → purine, pyrimidine → pyrimidine), so-called *transitions*,[8] are more likely than substitutions between nucleotide classes (purine ↔ pyrimidine), so-called *transversions*. For $w \to \infty$, the system 'forgets' its initial configuration as the result of the mixing caused by an increasing number of nucleotide substitutions (including backsubstitutions and multiple substitutions, as illustrated in Figure 1.4). Consequently, $P(y|x,w) \to \Pi(y)$ for $w \to \infty$, where $x, y \in \{A, C, G, T\}$, and $\Pi(y)$ is the equilibrium distribution (here: $\Pi(y) = 1/4 \;\forall\; y$).

Various nucleotide substitution models exist, depending on the number of free parameters, as I have reviewed in in Chapter 4 of [DH1]. For example, in the Kimura model, proposed in [75] and

[8] Unfortunately this terminology, which is used in molecular biology, leads to a certain ambiguity in the meaning of the word *transition*. When we talk about transitions between *states*, where the states are associated with nucleotides, as in Figure 1.10, a transition can be any nucleotide substitution event. When we talk about *transitions* as opposed to *transversions*, a transition refers to a certain type of nucleotide substitution.

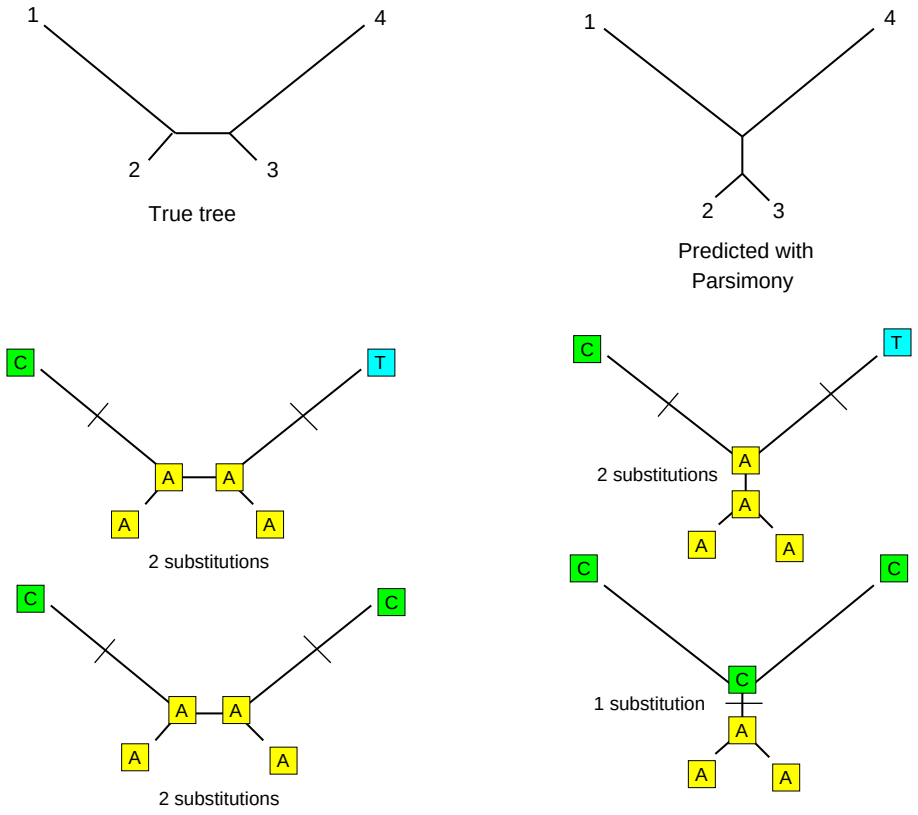


Figure 1.9: **Failure of parsimony.** The plot in the *top left* shows the true phylogenetic tree, in which the branch lengths of related taxa differ significantly. When the ratio of the long and short branch lengths exceeds a certain threshold, parsimony systematically predicts the wrong tree, shown in the *top right*. For an explanation, consider the tree in the *centre left*, which shows two nucleotide substitutions, one on either long branch. When these nucleotide substitutions are different, both tree topologies give the same score of 2 substitutions. Such nucleotide configurations are uninformative in that they do not prefer one tree over the other. When the nucleotide substitutions are identical, which happens, on average, in 25% of the cases, the tree on the right has a parsimony score of 1. This score is lower than that of the true tree, which is 2. Hence, such so-called homoplasious substitutions support the wrong tree. When the ratio of the external branch lengths exceeds a critical threshold, these "bad" substitutions outweigh, on average, the "good" substitutions that support the true tree. Consequently, parsimony will infer the wrong tree.

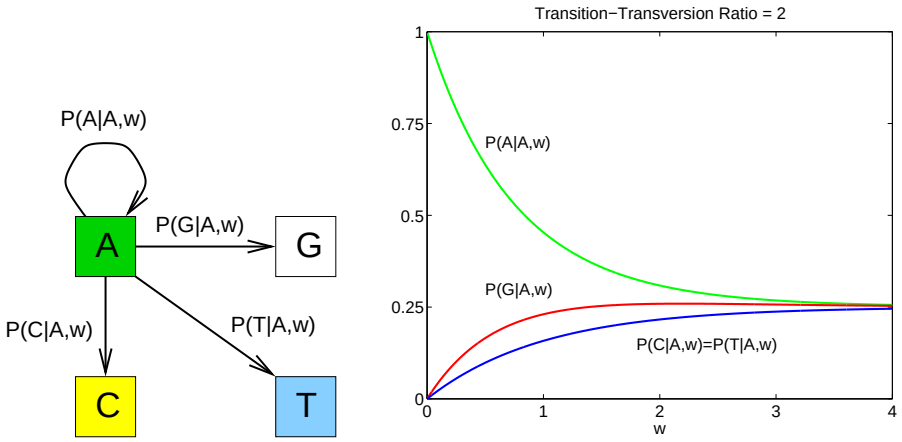


Figure 1.10: **Mathematical model of nucleotide substitution.** *Left:* Nucleotide substitutions are modelled as transitions in a 4-element state space. The transition probabilities depend on the phylogenetic time $w = \alpha t$, where t is physical time, and α is an unknown nucleotide substitution rate. *Right:* Dependence of the transition probabilities (vertical axis) on w (horizontal axis). The graphs were obtained from the Kimura model with a transition–transversion ratio of 2.

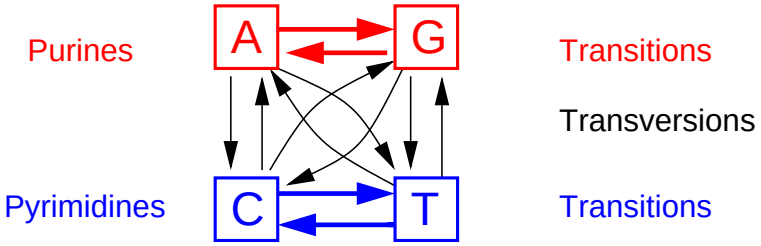


Figure 1.11: **Transitions versus transversions.** There are two types of nucleotides, denoted by purines and pyrimidines. Mutations within a group of nucleotides, called transitions and indicated here by the horizontal arrows, are more likely than mutations between groups of nucleotides. The latter are called transversions and are shown here by vertical arrows.

shown in Figure 1.12, the equilibrium distribution is uniform, and the transition-transversion ratio is a free parameter. If additionally the equilibrium probabilities are free parameters, as proposed in [58], the model is called the HKY85 model. Under certain regularity conditions, most notably the independence of nucleotide substitutions at different positions and the assumption that nucleotide processes are Markovian and homogeneous in time and space, we can compute the probability of a nucleotide configuration assigned to the leaves of a phylogenetic tree. I have provided a detailed exposition of this procedure in Chapter 4 of [DH1].

The upshot is that we can compute, for the tth column $\mathbf{y}_t$ in a DNA sequence alignment (meaning a vector with the nucleotides in the tth column), the probability $P(\mathbf{y}_t|\mathbf{w}, S)$. This probability depends on the tree topology, S, and the vector of branch lengths, $\mathbf{w}$, as illustrated in Figure 1.13. The respective computation can be repeated for every site, $1 \leq t \leq N$. Under the assumption that

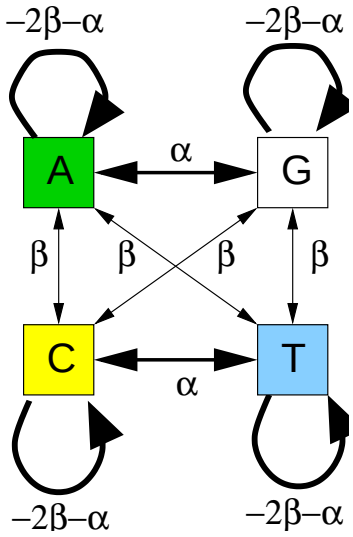


Figure 1.12: **Kimura model of nucleotide substitution.** The figure presents a graphical display of the Kimura model of nucleotide substitutions. The positive parameter α denotes the transition rate, while the positive parameter β denotes the transversion rate. The thickness of a line is related to the size of the respective rate parameter, indicating that transitions are more likely than transversions: $\alpha > \beta$.

mutations at different sites t are independent of each other – see above – the likelihood $P(\mathcal{D}|\mathbf{w}, S)$ of the whole DNA sequence alignment $\mathcal{D} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ factorizes:

$$P(\mathcal{D}|\mathbf{w}, S) \;=\; \prod_{t=1}^{N} P(\mathbf{y}_t|\mathbf{w}, S) \tag{1.2}$$

Equation (1.2) gives us an objective score or optimality function that opens the way to standard statistical inference: the frequentist approach based on maximum likelihood, and the Bayesian approach based on the posterior distribution of trees. Recall that for n species the number of unrooted phylogenetic trees is $(2n-5)!!$. Hence, the number of phylogenetic trees increases super-exponentially with the number of nodes, and the inference problem is NP-hard. In practice, heuristic (greedy) search procedures have to be adopted for finding the maximum likelihood configuration, and MCMC-based techniques are required for sampling trees from the posterior distribution in the Bayesian approach. An illustration is given in Figure 1.14. I have provided a detailed description of these methods in Chapter 4 of [DH1].

1.5 Bayesian versus Frequentist Inference

Given a data set $\mathcal{D}$ (in our case: a DNA sequence alignment), we want to learn the model S (in our case: a phylogenetic tree). There are two conceptually different approaches to this inference

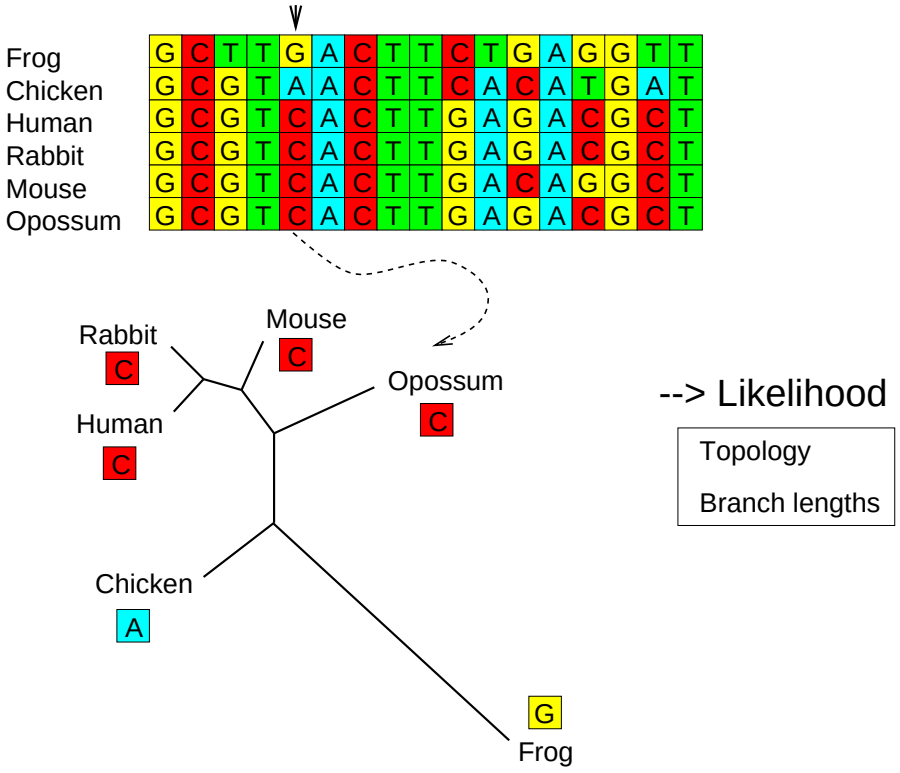


Figure 1.13: **Probabilistic approach to phylogenetics.** For a given column $\mathbf{y}_t$ in the DNA sequence alignment, a probability $P(\mathbf{y}_t|\mathbf{w}, S)$ can be computed, which depends on the tree topology, S, and the vector of branch lengths, $\mathbf{w}$. This can be done for every site, $1 \leq t \leq N$, which allows the computation of the likelihood $P(\mathcal{D}|\mathbf{w}, S)$ of the whole DNA sequence alignment $\mathcal{D} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$.

problem. In the frequentist or classical school of statistics, the aim is to find the 'best' model in a maximum likelihood sense:

$$S^*(\mathcal{D}) \;=\; \text{argmax}_S P(\mathcal{D}|S) \tag{1.3}$$

In order to assess the intrinsic uncertainty of inference, this estimation is conceived as being repeatable on fictitious data $\tilde{\mathcal{D}}$ which could have been generated in hypothetical parallel statistical universes governed by the same data generating process. From the distribution of $S^*(\tilde{\mathcal{D}})$ we can then obtain confidence intervals. In the Bayesian school of statistics, inference is based solely on the data $\mathcal{D}$ that have been observed. As opposed to the frequentist school, the model S is treated as a random variable, and inference is based on the posterior distribution $P(S|\mathcal{D})$. For instance, the best model can be found in a maximum a posteriori sense:

$$S^* \;=\; \text{argmax}_S P(S|\mathcal{D}) \tag{1.4}$$

There have been controversial debates about which school is 'right'. In the context of phylogenetics, frequentist statisticians argue that there is only one (unknown) phylogenetic tree, and its properties are fixed. On these grounds they object to the treatment of S as a random variable. However,

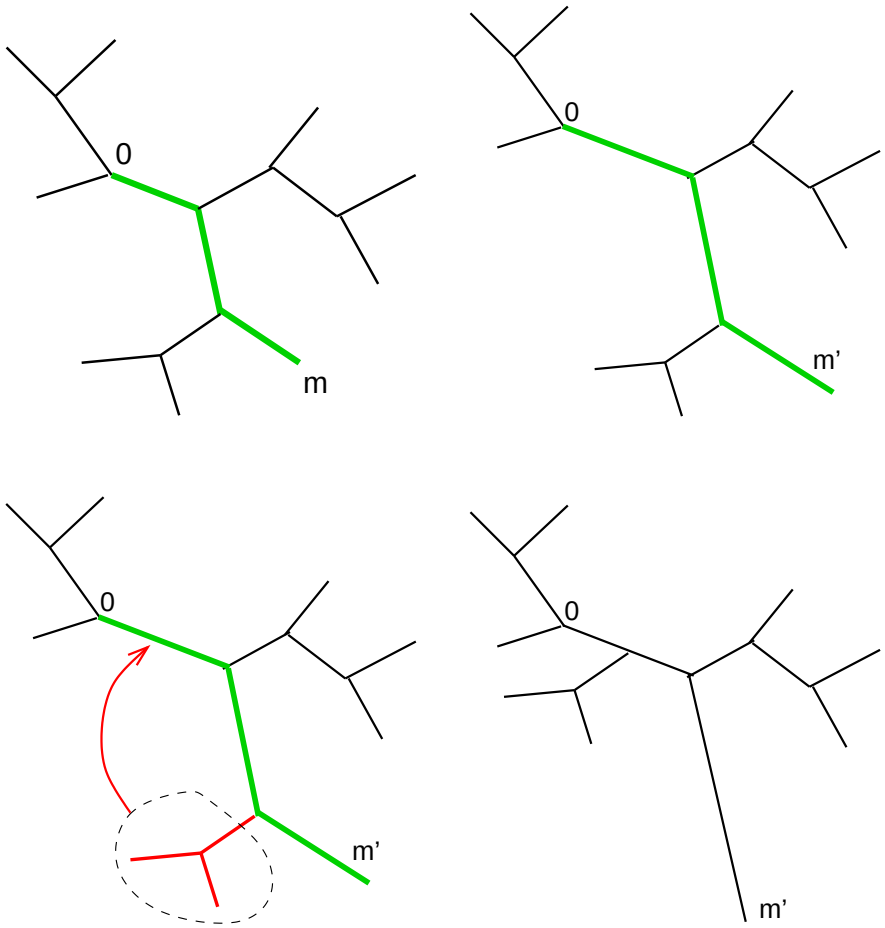


Figure 1.14: **MCMC/greedy search moves in tree space.** The figure illustrates a possible move in tree space, proposed in [80]. First, a backbone connecting four nodes is selected (*top left*) and extended or shrunk randomly (*top right*). Second, one of the two centre branches is randomly selected (*bottom left*) and then repositioned randomly on the backbone (*bottom right*). The first step changes the branch lengths $\mathbf{w}$. The second step may or may not change the tree topology S; in the present example, it does change the topology. In a greedy search, moves are accepted if and only if they increase the likelihood. In MCMC, moves are accepted according to the Metropolis-Hastings criterion [59].

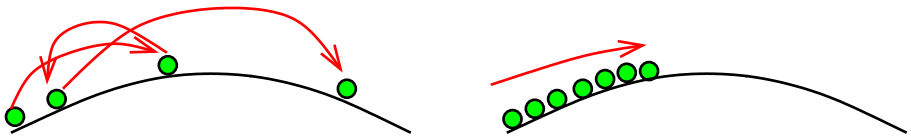


Figure 1.15: **A comparison between MCMC and bootstrapping.** The objective of both the Bayesian approach based on MCMC and the frequentist approach based on bootstrapping is to explore the uncertainty of inference, which is related to the curvature of the log likelihood surface. This figure gives an illustration for a simple one-dimensional case. The Bayesian MCMC approach, shown on the left, explores the log likelihood (or, in the case of a non-uniform prior, the log posterior) surface with a Markov chain. New configurations are first generated from a proposal distribution, and then accepted with a certain acceptance probability. Each accepted configuration is stored, and it contributes to the information about the curvature of the log likelihood surface. The frequentist approach, shown on the right, aims to find the maximum of the log likelihood surface. This maximum is a point estimate devoid of any information on uncertainty; consequently, this optimization is repeated several times for different bootstrap replicas. Information about estimation uncertainty is contained in the spread of the bootstrap maxima, that is, in the average deviation of the peak in response to resampling the data. Each optimization, however, is time-consuming, and all the information gathered along the trajectory leading to the peak of the log likelihood surface is eventually discarded. This waste of information renders bootstrapping much more computationally expensive than the Bayesian approach, where all the information along the MCMC trajectory is kept.

their reliance on a multitude of hypothetical parallel statistical universes brings about its own philosophical problems.

For a decision on which path to follow, note that for the type of complex problems typically encountered in computational biology, none of the two frameworks leads to any analytically tractable solutions. The standard approach pursued in frequentist statistics is based on bootstrapping [37]: data from hypothetical parallel statistical universes are approximated by sampling with replacement exemplars from the observed data, and then repeating the maximum likelihood estimation on these bootstrap replica. In the Bayesian framework, we typically address the intractability of the posterior distribution with Markov chain Monte Carlo (MCMC) [27]: we construct a Markov chain that converges to the posterior distribution in distribution, and then keep samples from that chain as an approximation to a sample from the true posterior distribution. Applied to the problem of phylogenetic inference, both approaches lead to a set of phylogenetic trees, from which confidence estimates can be computed. In particular, we are usually interested in the probability with which certain species are grouped together (clade support values). While in practice both the frequentist approach based on bootstrapping and the Bayesian approach based on MCMC tend to give similar results, the computational costs are drastically different. In a comparative evaluation on a DNA sequence alignment of 31 whale species, Larget and Simon [80] completed the Bayesian analysis in 7 hours of CPU time on a standard desktop PC. They estimated that repeating the analysis with the frequentist approach to the same degree of accuracy would have taken 175 days. This striking differences is caused by the fact that MCMC makes use of information gathered during its execution much more efficiently than bootstrapping, as illustrated in Figure 1.15. It is for this pragmatic rather than any dogmatic reason that in my own work I have elected to adopt the Bayesian approach.

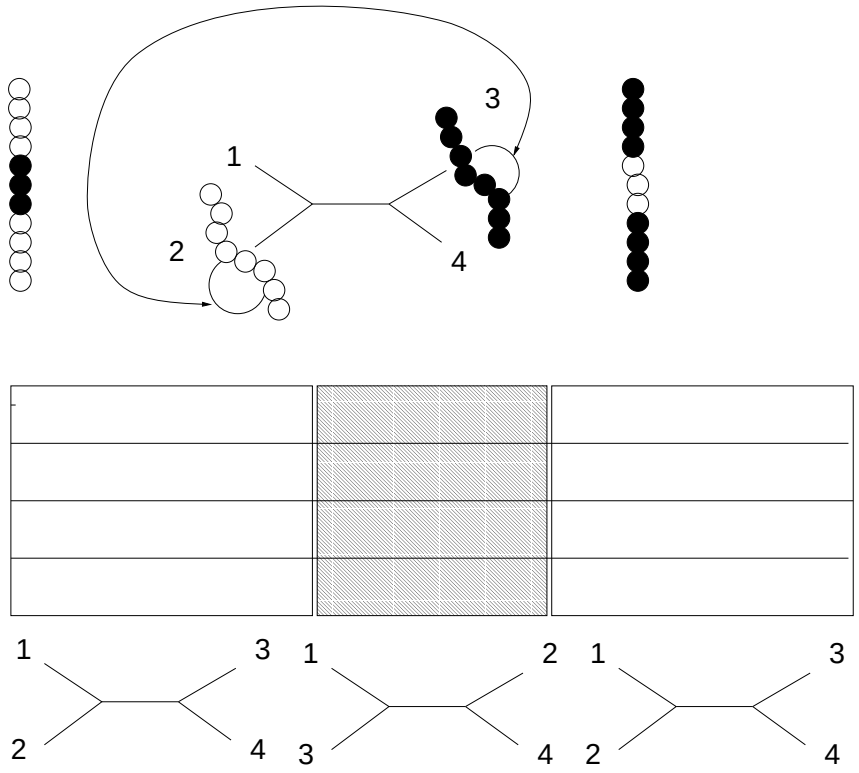


Figure 1.16: **Influence of recombination on phylogenetic inference.** The figure shows a hypothetical phylogenetic tree of four strains. Recombination is the transfer or exchange of DNA subsequences between different strains (top diagram, middle), which results in two so-called mosaic sequences (top diagram, margins). The affected region in the multiple DNA sequence alignment, shown by the shaded area in the middle diagram, seems to originate from a different tree topology, in which two branches of the phylogenetic tree have been swapped (bottom diagram, where the numbers at the leaves represent the four strains). A phylogenetic inference algorithm that does not identify this region is likely to give suboptimal results, since the estimation of the branch lengths of the predominant tree will be adversely affected by the conflicting signal coming from the recombinant region.

1.6 Recombination

The underlying assumption of the phylogenetic tree reconstruction methods reviewed above is that there is one set of hierarchical relationships among the taxa. While this is a reasonable approach when applied to most DNA sequence alignments, it can be violated in certain bacteria and viruses due to interspecific *recombination*. Recombination is a genetic process that results in the exchange or transfer of DNA/RNA subsequences and constitutes an important source of genetic diversification in certain bacteria and viruses. The resulting mixing of genetic material can lead to a change of the branching order (topology) of the phylogenetic tree in the affected region, which results in conflicting phylogenetic information from different regions of the alignment. Figure 1.16 demonstrates for a simple hypothetical scenario involving four strains how the transfer

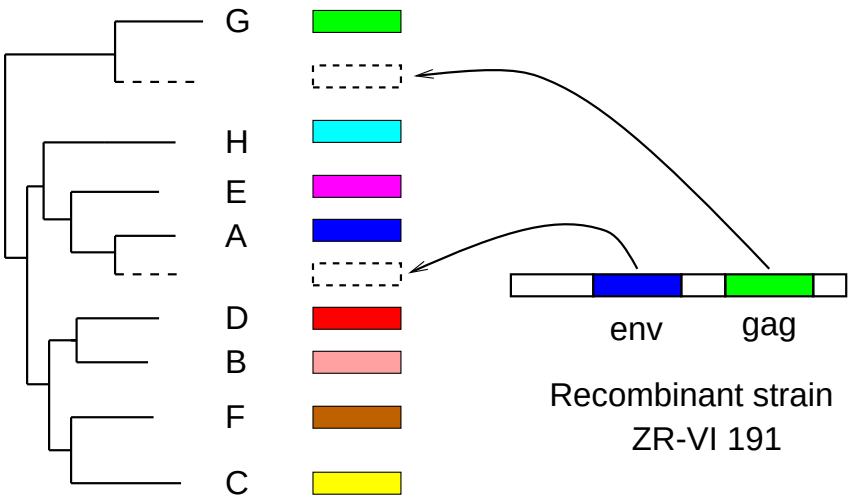


Figure 1.17: **Recombination in HIV-1.** The left diagram shows a phylogenetic tree of eight established subtypes of HIV-1. The right diagram represents the sequence of a recombinant strain (RS). When the phylogenetic analysis is based on the *env* gene, the RS strain is found to be most closely related to the A subtype. When the phylogenetic analysis is repeated for the *gag* gene, the RS strain seems to be most closely related to the G subtype. A conventional phylogenetic analysis treats the RS sequence as a monolithic entity. This fails to resolve the conflicting phylogenetic signals stemming from different regions of the alignment, and leads to a distorted "average" tree that is in a "limbo" state between the two true trees mentioned above. It is therefore vital to identify the mosaic structure of the RS strain, and to infer different phylogenetic trees for the different regions. Adapted from [116], by permission of Macmillan Magazines Limited.

or exchange of genetic material between different strains may lead to a change of the phylogenetic tree topology in the region affected by recombination, which results in conflicting phylogenetic information stemming from different regions of the alignment. Figure 1.17 illustrates the effects of recombination on the analysis of phylogenetic relationships among different HIV-1 strains. Both figures demonstrate that the presence of mosaic sequences resulting from recombination can lead to systematic errors in phylogenetic tree estimation. Their detection, therefore, is a crucial prerequisite for consistently inferring the evolutionary history of a set of DNA/RNA sequences. Additionally, the recent advent of multiple-resistant pathogens has led to an increased interest in interspecific recombination as an important, and previously underestimated, source of genetic diversification in bacteria and viruses. The discovery of a surprisingly high frequency of mosaic RNA sequences in HIV-1 suggests that a substantial proportion of AIDS patients have been coinfected with HIV-1 strains belonging to different subtypes, and that recombination between these genomes can occur *in vivo* to generate new biologically active viruses [116]. A phylogenetic analysis of the bacterial genera *Neisseria* and *Streptococcus* has revealed that the introduction of blocks of DNA from penicillin-resistant non-pathogenic strains into sensitive pathogenic strains has led to new strains that are both pathogenic *and* resistant [92]. Thus the presence of interspecific recombination raises the possibility that bacteria and viruses can acquire biologically important traits through the exchange and transfer of genetic material.

1.7 Traditional Methods for Detecting Recomination

Several methods for detecting interspecific recombination have been developed – following up on the seminal paper by Maynard Smith [92] – and I have reviewed the classical approaches in [DH1], Chapter 5. None of these methods is without shortcomings. While the reticulated structure of *phylogenetic networks* [7, 136, 137] indicates the presence of recombination, it does not easily allow the location and the changepoints of the recombinant regions in the DNA sequence alignment. By discarding all but the polymorphic sites in the DNA sequence alignment and ignoring explicit phylogenetic relationships, the method of *maximum chi-square* [92] does not make the most efficient use of the information contained in the DNA sequence alignment and, consequently, detects the location of recombinant regions at comparatively poor resolution. The method of *Partial Likelihoods Assessed Through Optimization (PLATO)*, proposed in [53], becomes increasingly unreliable as the recombinant regions grow in length. The DSS statistic of TOPAL [93, 94] is computed from pairwise distances between DNA sequences, which suffers from the intrinsic information loss discussed in Section 1.2 and illustrated in Figure 1.5. Finally, methods based on maximum parsimony, following up on Hein's seminal RECPARS algorithm [63], are inflicted by the inconsistency problem described in Section 1.3. They also depend on various tuning parameters, which cannot be inferred from the data and, hence, have to be heuristically set by the user.

1.8 Methodological Improvements

In the next three chapters, I will review the methods for detecting recombination and mosaic structures in DNA sequence alignments that I have developed and co-developed. These methods are based on phylogenetic trees inferred from the DNA sequence alignment with the likelihood paradigm. As opposed to the DSS statistic of TOPAL and the method of maximum chi-square, they avoid any intrinsic loss of information from the data. As opposed to RECPARS, they avoid the inconsistency problem inherent in maximum parsimony. As opposed to phylogenetic networks, they allow the location of recombinant segments to be inferred. The more advanced models described in Chapters 3 and 4 are based on Bayesian hierarchical models, which allow all parameters to be consistently inferred from the data.

Chapter 2

Detecting Recombination with a Sliding Window Method

Summarizing my work in [DH11,DH17,DH21,DH22,DH29], this chapter describes a graphical exploratory method based on moving a window along the DNA sequence alignment and computing a divergence measure of a phylogenetic signal as a potential indicatior of evidence for recombination. The phylogenetic signal chosen is the posterior distribution of phylogenetic tree topologies, which is based on the Bayesian approach to phylogenetic inference, and which is practically computed by repeatedly running Markov chain Monte Carlo (MCMC) simulations on different subsets of the DNA sequence alignment [DH29]. This method can be refined with a topology based pruning scheme [DH21] and the combination with a Bayesian hidden Markov model [DH17]. The algorithms have been implemented in a user-friendly software package [DH11,DH22].

2.1 Method A: Probabilistic Divergence Method (PDM)

This section reviews my work with Frank Wright, published in [DH29]. The idea of window-based detection methods is to slide a window along the DNA sequence alignment and monitor changes in the posterior distribution of the phylogenetic tree topology. An illustration of the concept is given in Figures 2.1–2.3. For a formal introduction, consider a given alignment of DNA sequences, $\mathcal{D}$, from which we select a consecutive subset $\mathcal{D}_t$ of predefined width W, centred on the tth site of the

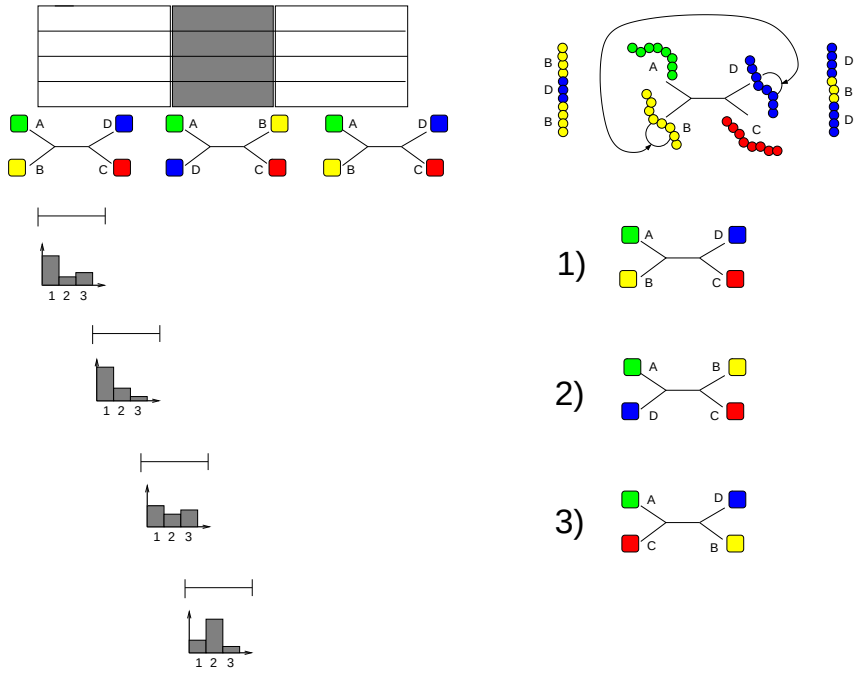


Figure 2.1: **Illustration of the probabilistic divergence method (PDM).** The *top right* subfigure shows a recombination scenario in a four-species tree similar to Figure 1.16. The *top left* subfigure shows the resulting sequence alignment, where the tree topology in the centre block is different from the topology in the flanking regions as a consequence of recombination. The subfigure in the *bottom right* shows the three possible tree topologies. The *bottom left* subfigure shows the posterior distribution over tree topologies conditional on different subregions of the alignment, selected by a moving window. When moving the window into a region that corresponds to a different tree topology, the posterior distribution can be expected to change markedly.

alignment. Let S be an integer label for tree topologies, and define

$$P_S(t) \;:=\; P(S|\mathcal{D}_t) \;=\; \int\int P(S,\mathbf{w},\boldsymbol{\theta}|\mathcal{D}_t)d\mathbf{w}d\boldsymbol{\theta} \tag{2.1}$$

This is the marginal posterior distribution of tree topologies S, conditional on the "window" $\mathcal{D}_t$, which includes a marginalization over the branch lengths $\mathbf{w}$ (see Figure 1.13) and the parameters of the nucleotide substitution model $\boldsymbol{\theta}$ (e.g. in Figure 1.12 this refers to the transition-transversion ratio α/β). In practice the integral in (2.1) is solved numerically by means of a Markov chain Monte Carlo (MCMC) simulation, and can be carried out with standard Bayesian phylogenetic inference programs, like BAMBE [80]. This MCMC simulation yields a sample of triples $\{S_{ti},\mathbf{w}_{ti},\boldsymbol{\theta}_{ti}\}_{i=1}^{M}$ simulated from the joint posterior distribution $P(S,\mathbf{w},\boldsymbol{\theta}|\mathcal{D}_t)$. We then replace the true posterior distribution by the empirical distribution

$$P(S,\mathbf{w},\boldsymbol{\theta}|\mathcal{D}_t) \quad\approx\quad \frac{1}{M}\sum_{i=1}^{M}\delta_{S,S_{ti}}\delta(\mathbf{w}-\mathbf{w}_{ti})\delta(\boldsymbol{\theta}-\boldsymbol{\theta}_{ti}) \tag{2.2}$$

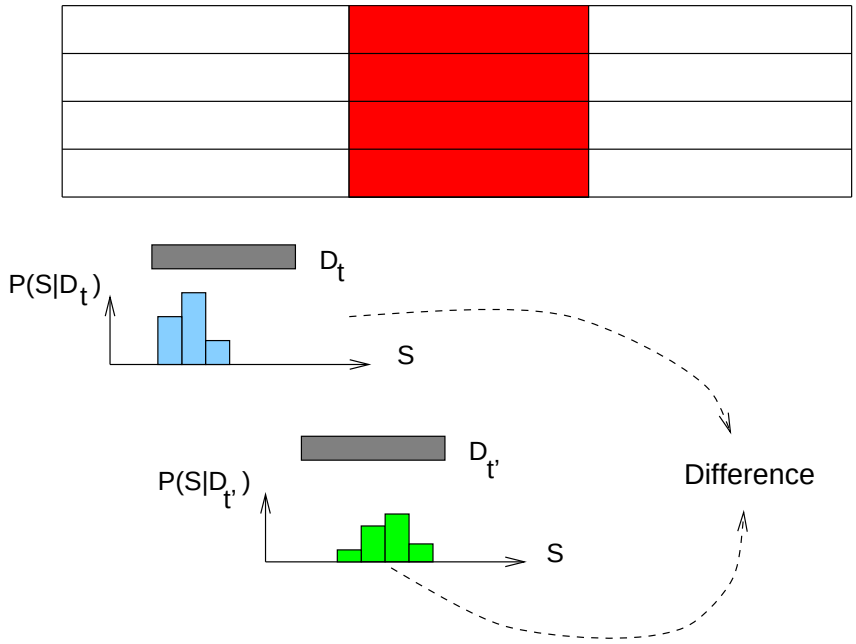


Figure 2.2: **Detecting recombination with the PDM method.** The figure shows the posterior distribution $P(S|D_t)$ of tree topologies S conditional on two subsets D_t and $D_{t'}$ selected by a moving window. When the window is moved into a recombinant region, the posterior distribution $P(S|D_t)$ can be expected to change significantly, and therefore to lead to a high PDM score.

where $\delta_{S,S_{ti}}$ denotes the Kronecker delta symbol, and $\delta(.)$ is the delta function, which is equal to 1 if $S_t = S_{t-1}$ and 0 otherwise. Inserting (2.2) into (2.1) gives:

$$P_S(t) \;=\; \frac{1}{M}\sum_{i=1}^{M}\delta_{S,S_{ti}} \;=\; \frac{M_S(t)}{M} \tag{2.3}$$

where $M_S(t)$ denotes the number of times a tree has been found to have topology S.

The basic idea of the probabilistic divergence method (PDM) for detecting recombinant regions is to move the window $\mathcal{D}_t$ along the alignment and to monitor the distribution $P_S(t)$. We would expect a substantial change in the shape of this distribution as we move the window into a recombinant region, as illustrated in Figures 2.1 and 2.2. The question, then, is how to easily monitor such a change and how to estimate its significance. To this end we consider the Kullback-Leibler divergence as the natural distance measure in probability space:

$$K(P,Q) \;=\; \sum_{S} P_S \log\left(\frac{P_S}{Q_S}\right) \tag{2.4}$$

in which P and Q denote probability distributions. A standard result from information theory is the non-negativity of the Kullback-Leibler divergence (see e.g. Chapter 2 in [DH1] and [104]), i.e.

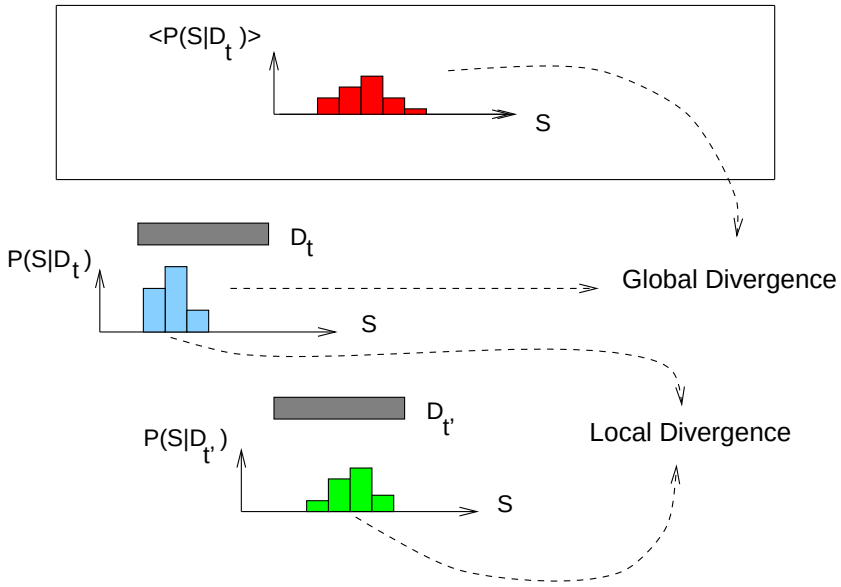


Figure 2.3: **Global and local PDM.** The figure shows the posterior distribution $P(S|D_t)$ of tree topologies S conditional on two subsets D_t and $D_{t'}$ selected by a moving window. The *local PDM* compares two distributions conditional on adjacent windows. This measure is a local divergence score. Alternatively, one can first obtain a reference distribution $\langle P(S|D_t)\rangle$ by averaging $P(S|D_t)$ over all window positions, and then compute a *global PDM* between $\langle P(S|D_t)\rangle$ and each local distribution $P(S|D_t)$.

$K(P,Q)$ is always non-negative, and zero if and only if $P = Q$. To estimate the divergence between the local and the global distribution, define P as in (2.3) and Q as the average distribution:

$$d[P(t),\overline{P}] \;=\; K[P(t),\overline{P}]; \qquad \overline{P}_S \;=\; \frac{1}{M_w}\sum_{t=1}^{M_w} P_S(t) \tag{2.5}$$

where M_w is the number of different window positions. This divergence score is the *global PDM*. Since $\mathrm{Support}(P) \subseteq \mathrm{Support}(\overline{P})$, where $\mathrm{Support}(P)$ is the set of all topologies S for which $P_S \neq 0$, the non-singularity of the expression on the left-hand side of (2.5) is guaranteed. To determine the divergence between two local distributions $P(t)$ and $P(t+\Delta t)$ – conditional on two adjacent windows with centre positions t and $t+\Delta t$ – a slightly modified divergence measure[1] is used:

$$\begin{aligned} &d[P(t),P(t+\Delta t)] = \\ &\quad\frac{1}{2}\left[K\left(P(t),\frac{P(t)+P(t+\Delta t)}{2}\right)+K\left(P(t+\Delta t),\frac{P(t)+P(t+\Delta t)}{2}\right)\right] \end{aligned} \tag{2.6}$$

Note again that $\mathrm{Support}[P(t)],\mathrm{Support}[P(t+\Delta t)] \subseteq \mathrm{Support}\left[\frac{P(t)+P(t+\Delta t)}{2}\right]$ guarantees the non-singularity of $d[P(t),P(t+\Delta t)]$. To estimate whether the observed divergence measures are significantly different from zero, we can use the fact that under the null hypothesis of no recombination,

[1]The divergence measure of (2.6) was introduced by Sibson; see, for instance, [78], Chapter 14.

$P = Q$, the Kullback-Leibler divergence is asymptotically chi-squared distributed [66]. A better yet computationally more expensive approach is based on parametric bootstrapping. First, a phylogenetic tree is estimated from the whole sequence alignment under the assumption of no recombination. Second, DNA sequence alignments without recombination are generated from this tree. Third, the method is repeatedly applied to these alignments. In this way we obtain a distribution of PDM scores under the null hypothesis of no recombination, from which p-values and significance indicators can be computed.

The local PDM (2.6) depends on the distance between two windows, Δt. It seems natural to choose two consecutive windows. However, by allowing a certain overlap between the windows the spatial resolution of the detection method can be improved. In the empirical study described in [DH29] I found that the best results can be obtained by averaging over different degrees of window overlap:

$$\overline{d} \;=\; \frac{1}{A}\sum_{a=1}^{A} d[P(t), P(t + a\Delta t)] \tag{2.7}$$

where $d(.)$ is defined in (2.6), and the average is over all window overlaps between 50% and 90%.

I have assessed the performance of the PDM method on various synthetic and real-world DNA sequence alignments subject to recombination; see [DH29] for details. A comparative evaluation revealed that PDM detects recombinant regions with a higher accuracy than the traditional methods reviewed in Section 1.7. In particular it improves on [93, 94] in three aspects. First, by using a likelihood score, the information loss inherent in a score based on pairwise sequence distances, as discussed in Section 1.2 and illustrated in Figure 1.5, is prevented. Second, a single optimized reference tree is replaced by a distribution over trees, which captures the intrinsic uncertainty of tree estimation from short sequence alignments. Third, the method focuses on topology changes and thereby avoids the potentially confounding effect of rate heterogeneity, which I will discuss in more detail in Chapter 4.

The method has been implemented in a user-friendly software package [DH22], and in a joint project with Miles Armstrong, Mark Phillips and Vivian Bloks I have successfully applied it to the detection of recombination in an alignment of mitochondrial DNA sequence from various populations of the potato cyst nematode *Globodera pallida* [15].

2.2 Method B: Improvement by Pruning

In this section I summarize the work described in [DH21]. The principal shortcoming of the PDM method is illustrated in Figure 2.4. A substantial change in the tree topology caused by a recombination event is not distinguished from changes in the clade configurations of closely related

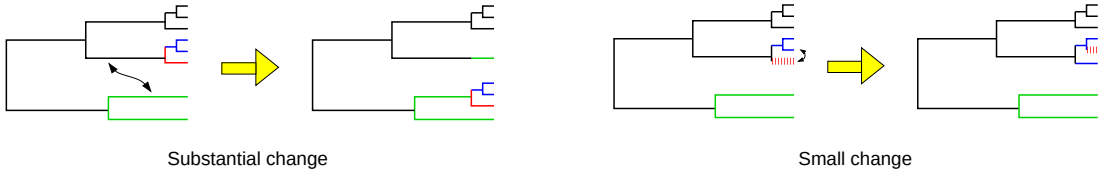


Figure 2.4: **Shortcoming of the PDM method.** A substantial change in the branching structure of a phylogenetic tree, illustrated on the left, is indicative of a recombination event. Changes of the branching structure that only involve closely related strains, shown on the right, may result as a mere consequence of statistical fluctuations. The PDM method described in the previous section does not distinguish between these two types of changes in the tree topology, which renders the approach suboptimal.

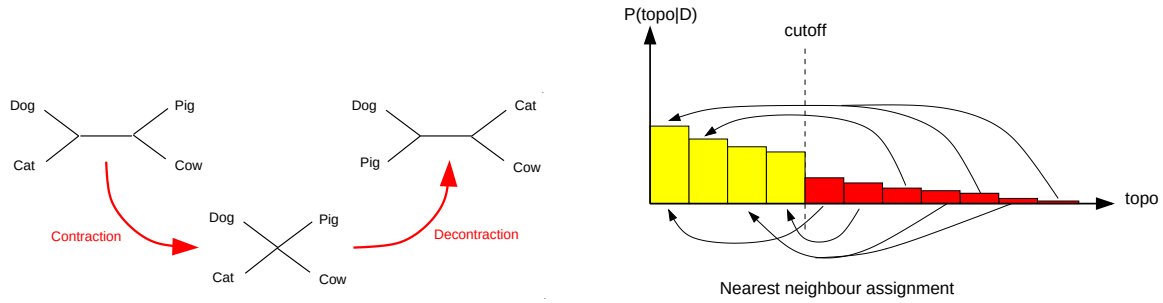


Figure 2.5: **RF-distance based pruning.** *Left panel:* The Robinson-Foulds (RF) distance between tree topologies is defined as the minimum number of contraction and decontraction operations needed to transform one topology into the other. *Right panel:* On the average posterior distribution of tree topologies, averaged over all sliding window positions, a cutoff threshold is defined. Tree topologies above this threshold are kept as "principal topologies". Tree topologies below the threshold are assigned to the principal topology with the minimum RF distance. After this re-assignment, the posterior distribution is re-normalized.

taxa. As the number of sequences in the alignment increases, so does the number of such within-clade reconfigurations. This is because for an increased number of taxa the posterior distribution over tree topologies, $P(S|\mathcal{D}_t)$, becomes increasingly disperse unless the size of the data set $\mathcal{D}_t$ is increased. An increased amount of data $\mathcal{D}_t$, however, corresponds to an increased length of the sliding window, which compromises the spatial resolution of the detection and is not an option for short alignments.

A simple pruning scheme

A possible remedy for this "dispersion" problem is to include information on the amount of change between different tree topologies and thereby distinguish between the scenarios depicted in Figure 2.4. Since branch lengths have been marginalized over to avoid the confounding effect from rate heterogeneity[2], the difference has to be estimated solely on the basis of topological information. A possible metric in the space of tree topologies was introduced in [117]. The idea is illustrated in Figure 2.5. Consider the two complimentary operations of merging two existing nodes into one, and

[2]Rate heterogeneity will be discussed in Section 4.

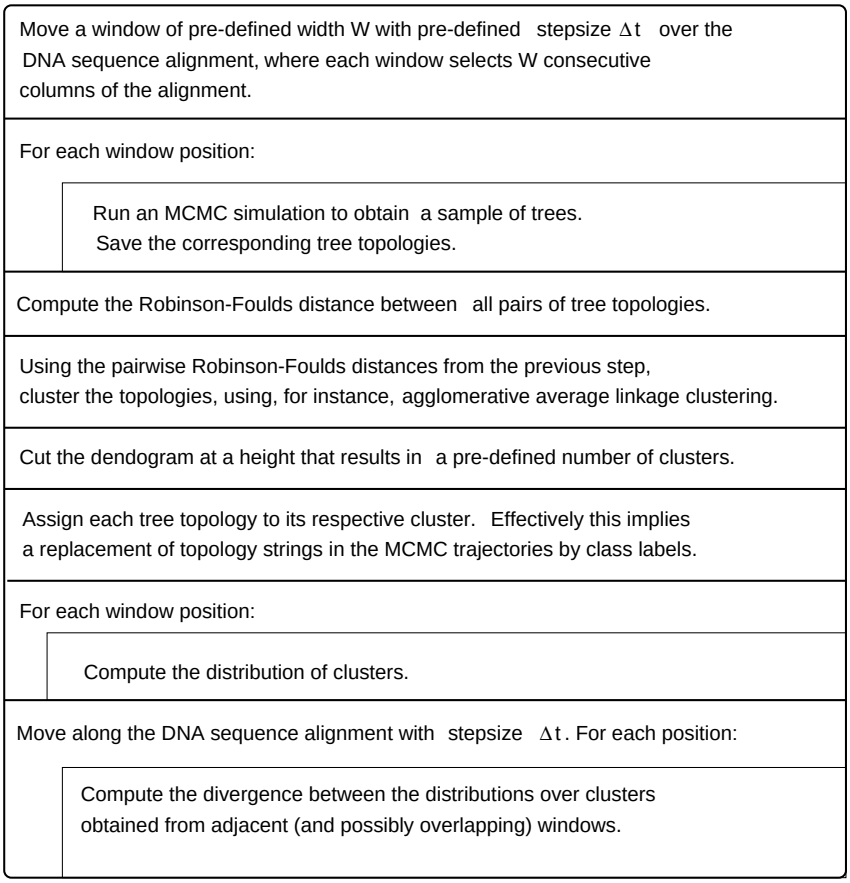


Figure 2.6: **Pseudo code of the pruning by clustering algorithm.** The figure summarizes the various steps involved in the computation of the pruned probabilistic divergence signal.

splitting an existing node into two. The Robinson-Foulds (RF) distance between tree topologies S_1 and S_2 is defined as the minimum number of operations required to transform S_1 into S_2 (or S_2 into S_1). Robinson and Foulds [117] derived a simple algorithm for the practical computation, which has been implemented in various phylogeny software packages, like PHYLIP [42]. Define $\langle P(S|\mathcal{D})\rangle$ to denote the average posterior distribution of tree topologies, averaged over all sliding window positions:

$$\langle P(S|\mathcal{D})\rangle \;=\; \frac{1}{N_W}\sum_{t=1}^{N_W} P(S|\mathcal{D}_t) \tag{2.8}$$

in which N_W denotes the total number of window positions. The support of the average posterior distribution is the set of those tree topologies for which $\langle P(S|\mathcal{D})\rangle \neq 0$, that is, the set of all tree topologies visited during the MCMC simulations. In an application of the PDM method to a DNA sequence alignment of ten strains of Hepatitis-B virus, I found a support of 126 distinct tree topologies. Usually, we do not expect to find over hundred recombination events in an alignment of

about 3000 nucleotides. Hence most of topology changes correspond to within-clade reconfigurations resulting from diffuse posterior distributions, which degrades the reliability and efficacy of the PDM method.

A way to proceed is to reduce the dispersion of the posterior distribution by reducing the cardinality of the support of $\langle P(S|\mathcal{D})\rangle$. This can be effected with a pruning scheme based on the RF distance. First, identify a set of principal tree topologies, for instance, those that maximize $\langle P(S|\mathcal{D})\rangle$. Next, assign each non-principal tree topology to the principal topology with the minimum RF distance. Finally, renormalize the posterior distributions $P(S|\mathcal{D}_t)$ and recompute the PDM signal. An illustration is given in Figure 2.5. The reduction in the dispersion of $P(S|\mathcal{D}_t)$ is likely to reduce the noise in the PDM signal. Note that similar pruning methods are used in machine learning to improve the generalization performance of a predictor [13]. Also note that the pruning of the support of $\langle P(S|\mathcal{D})\rangle$ has some similarity with a Bayesian approach in that it brings the data-based prediction in line with our prior assumption about the expected number of recombination events.

Improved pruning by clustering

The pruning scheme described in the previous subsection has an obvious disadvantage: when the recombinant region is short, the set of principal topologies may not contain any topology that reflects the recombination event. Also, the assignment of non-principal to principal topologies can be interpreted as the first step of an incomplete K-means clustering procedure. Such topology-based clustering of phylogenetic trees was proposed in [135] to reduce the information loss incurred by a single consensus tree approach. It here offers the obvious method to be used for pruning. Based on the RF distance, the MCMC sample of tree topologies, $\{S_1, \ldots, S_M\}$, is clustered. For a given number of clusters K, tree topologies are assigned to their respective cluster $\mathcal{C}_1, \ldots, \mathcal{C}_K$:

$$I(S_i \in \mathcal{C}_k) \;=\; \left[\begin{array}{l} 1 \text{ if } S_i \in \mathcal{C}_k \\ 0 \text{ if } S_i \notin \mathcal{C}_k \end{array} \right. \tag{2.9}$$

Now, define the posterior distribution over clusters,

$$P(\mathcal{C}_k|\mathcal{D}_t) \;=\; \sum_S I(S \in \mathcal{C}_k) P(S|\mathcal{D}_t) \tag{2.10}$$

which is computed from the MCMC sample of tree topologies, $\{S_1, \ldots, S_M\}$:

$$P(\mathcal{C}_k|\mathcal{D}_t) \;=\; \frac{1}{M} \sum_{i=1}^{M} I(S_i \in \mathcal{C}_k) \tag{2.11}$$

An improved pruning scheme can be achieved by using the posterior distribution over classes (2.11) instead of the original posterior distribution over topologies (2.3) in the computation of the divergence measure $D(t)$ in (2.6). I will henceforth refer to this novel approach of computing $D(t)$ from $P(\mathcal{C}_k|\mathcal{D}_t)$ rather than $P(S|\mathcal{D}_t)$ as the *pruned PDM* method.

Note that while this method addresses the shortcomings of the original PDM method and the simple pruning scheme of Figure 2.5, it is still heuristic in that the choice of the clustering algorithm is arbitrary. Our work in [DH17] is based on the findings in [135]. When applying MCMC to sample trees from the posterior distribution, as described in [80], one is faced with the problem of summarizing the information contained in the MCMC sample succinctly. A widely applied method is to resolve the conflicts within the obtained sample of tree topologies by computing a consensus tree, which, however, may incur a substantial information loss. Stockham et al. [135] investigated a post-processing alternative, by which trees are first divided into subsets with some clustering algorithm based on the RF distance, and each cluster is then characterized by its own consensus tree. The authors assessed various clustering algorithms with different measures of information loss. They found that bottom-up average linkage clustering outperformed top-down K-means clustering and the method of phylogenetic islands [89] in terms of complexity versus information content, and we therefore used the former scheme in our work [DH21].

The pruning-by-clustering procedure we proposed in [DH21] thus works as follows. First, generate an exhaustive list of all distinct tree topologies sampled in the complete set of MCMC simulations, that is, the MCMC simulations carried out for all window positions. Next, compute all pairwise RF distances between topologies, and infer a dendrogram of topologies with agglomerative average linkage clustering from these distances. Cut this dendrogram at a certain height such that it disintegrates into a pre-defined number of clusters. Finally, assign each tree topology to its respective cluster and use this assignment to compute the posterior distribution over clusters by application of (2.10). This posterior distribution over clusters supersedes the original posterior distribution over topologies in the computation of the divergence measure (2.6). A summary of the proposed algorithm is given in Figure 2.6.

Findings

Figure 2.7 illustrates the effect of the pruning scheme. I simulated two recombination events along with a differently diverged region (which has a confounding effect, see Chapter 4) in a synthetic DNA sequence alignment of 8 taxa with 6000 nucleotides each. There are four recombination changepoints, delimiting two recombinant regions of length 500 nucleotides. When applying the unpruned PDM method discussed in the previous section, decreasing the window size deteriorates the detection accuracy: the posterior distribution over tree topologies becomes more diffuse, leading to a more erratic PDM signal and the identification of spurious changepoints. This problem is avoided with the proposed pruning scheme, which in combination with the decreased window size results in an improved spatial resolution. Details of the simulations as well as a discussion of several other comparative evaluations on both synthetic and real DNA sequence alignments can be found in [DH21].

The method has been implemented in a user-friendly software package [DH11].

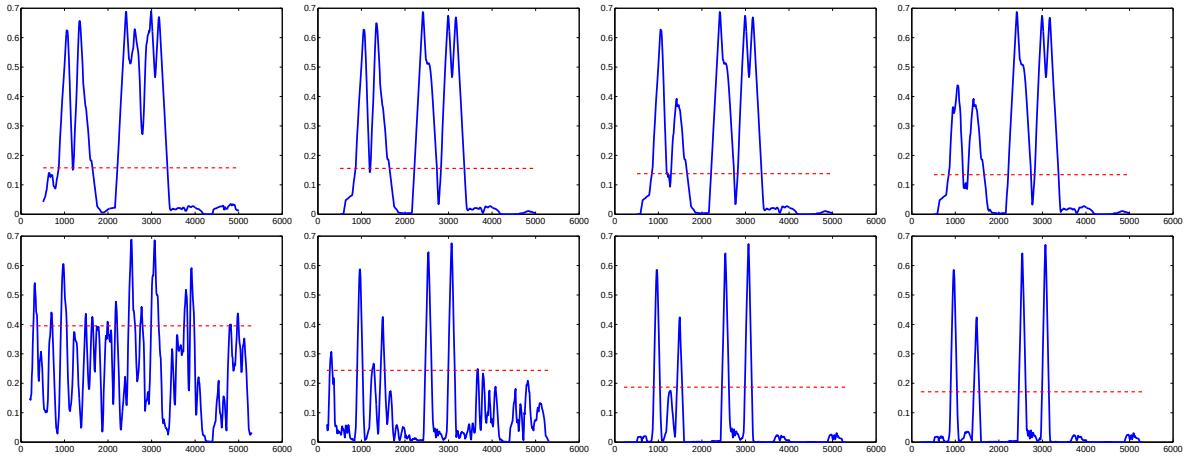


Figure 2.7: **Detection of recombination in a synthetic DNA sequence alignment with the pruned PDM method.** The graphs show probabilistic divergence signals using a window size of 500 (top panels) and 200 (bottom panels) nucleotides. The dashed horizontal lines show the 99 percentiles under the null hypothesis of no recombination. From left to right: no pruning (resulting in 50 tree topologies when the window size is 500, and 406 topologies for a window size of 200), pruning down to 7, 5, and 3 clusters. The true mosaic structure of the sequence alignment is as follows. Ancient recombination event: sites 1000–1500; recent recombination event: sites 2500–3000; differently diverged region: sites 4000–4500. Note that the latter has a confounding effect, and should be avoided by successful detection methods.

2.3 Method C: Combination with Hidden Markov Models (HMMs)

This section reviews my work with Anna Kedzierska [DH17], which aimed at combining the window-based detection methods from the previous sections with hidden Markov models (HMMs). HMMs provide a powerful tool widely used in bioinformatics [6], and they have been successfully applied to the segmentation of DNA sequences [19]. Here, the objective is to locate homogeneous segments that are compositionally different from the rest of the sequence. The hidden states represent the homogeneous segments to be detected, which are characterized by their distribution of nucleotides, or by their first-order Markovian transition probabilities between nucleotides [19]. A critical question is to infer how many different segment types a DNA sequence is composed of. To this end, Boys et al. [17, 18] adopted a Bayesian approach and sampled the number of hidden states from the respective posterior distribution with reversible jump (RJ) Markov chain Monte Carlo (MCMC).

The problem of detecting recombination is related to the segmentation problem described above, but differs from it in two important aspects. First, the data to be segmented is not a single DNA sequence, but an alignment of several DNA sequences. Second, homogeneity in a segment is not defined with respect to the nucleotide composition, but with respect to the underlying phylogenetic tree topology. A possible approach, proposed in the next chapter, is to associate the hidden

states with individual tree topologies. However, the number of tree topologies increases super-exponentially with the number of sequences in the alignment, and this method is therefore restricted to small numbers of taxa. The method we proposed in [DH17] is based on the idea of combining an HMM with the window-based detection method of the previous sections, Sections 2.1 and 2.2. First, we cut the DNA sequence alignment into segments of fixed length. For each of these segments, we sample phylogenetic trees from the posterior distribution – conditional on the respective segment – with MCMC. Second, we apply a clustering algorithm based on an appropriate distance metric in tree topology space, as described in the previous section, and we convert the posterior distributions over tree topologies into posterior distributions over clusters. Third, we associate a position in the alignment with the corresponding segment centre and characterize it by the respective posterior distribution over topology clusters. These posterior distributions summarize the information in the DNA sequence alignment and correspond to the observation of individual nucleotides in a single DNA sequence. Fourth, we apply a Bayesian hidden Markov model. We infer the optimal number of homogeneous segments, that is, the optimal number of hidden states, with RJMCMC, as in [17, 18]. For the actual location of the recombination changepoints, we compute the marginal posterior distributions over the hidden states, where state transitions indicate changepoints between putative mosaic segments in the DNA sequence alignment. An illustration of the scheme is given in Figure 2.8, and a pseudo code description of the algorithm is available from Figure 2.9.

Hence, our method described in [DH17] is effectively a combination of various modelling and inference schemes: the Bayesian approach to phylogenetics with MCMC [80], the probabilistic divergence measures introduced in [DH21,DH29] and described in Sections 2.1 and 2.2, the clustering of tree topologies [135] based on an appropriate distance metric [117], the segmentation of DNA sequences with HMMs [19], and the Bayesian approach to inference in HMMs with RJMCMC [17, 18, 115].

Our method is, in essence, a modification of the method described in the previous section. It utilizes, in fact, the same pre-processing scheme: First, we run several MCMC simulations on relatively short DNA sequence alignment segments. Next, we condense the complexity of the resulting topology space with a clustering method based on a topological distance measure. The main difference between the method described here and the one from Section 2.2 consists in the way the resulting position-dependent posterior distributions over topology clusters are exploited. The approach in [DH21], described in the previous section, consists in computing a probabilistic divergence measure, where 'significantly large' peaks indicate putative recombination changepoints. The difficulty, then, is to decide when a peak can be classified as 'significantly large'. This question can be decided from the distribution of peak heights under the null hypothesis of no recombination. This distribution can, in principle, be computed with a bootstrapping scheme. However, bootstrapping requires us to re-run the phylogenetic pre-processing step on hundreds of bootstrap replicas. This leads to very high computational costs, as we have to run new MCMC simulations on each bootstrap replica. Alternatively, one could resort to asymptotic and approximate methods; however, this inevitably compromises the estimation accuracy. The method described in the present section addresses the problem of identifying significant mosaic segments by formulating it as a model selection problem

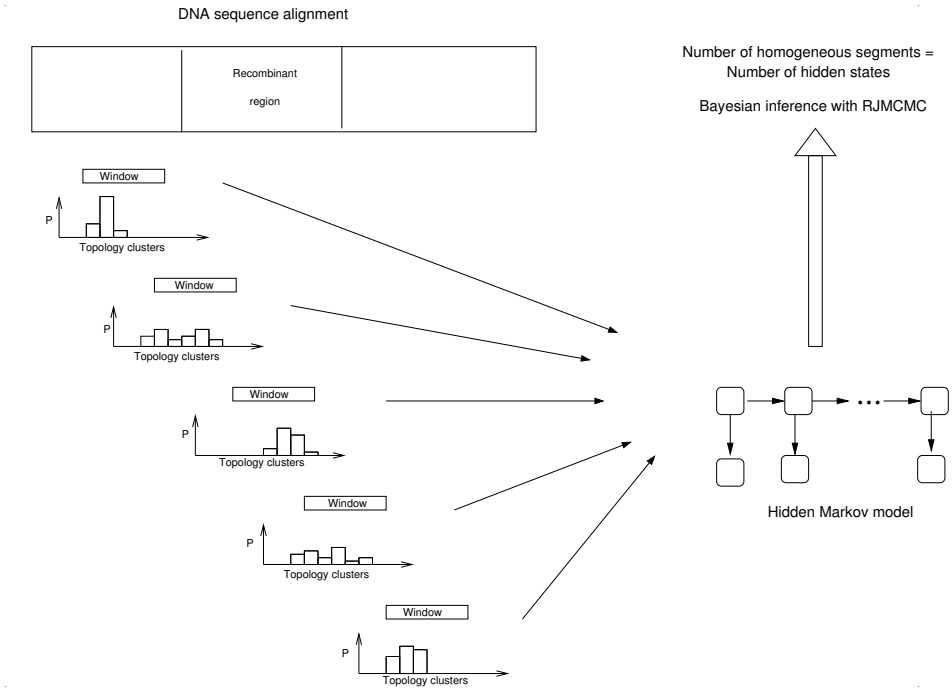


Figure 2.8: **Illustration of the combination with HMMs.** Left: A sliding window is moved along the DNA sequence alignment, which may contain a recombinant region (shown in the centre). For each window position, we obtain a distribution of tree topologies with MCMC. This distribution over tree topologies can be converted into a distribution over topology clusters, as in Figure 2.6. Right: The distributions thus obtained form the emission probabilities of a hidden Markov model (HMM). The hidden states of the HMM represent homogeneous segments of the DNA sequence alignment. Model order and parameters of the HMM are inferred in a Bayesian sense with RJMCMC. In the present example, we would expect to find two distinct hidden states corresponding to the recombinant region and the flanking regions in the DNA sequence alignment.

in Bayesian HMMs. This is closely related to the DNA sequence segmentation with HMMs [19], where each hidden state represents a homogeneous DNA segment with a characteristic distribution over nucleotides, and the true number of homogeneous DNA segments has to be determined. Boys et al. [17, 18] address this problem by turning the number of hidden states into a random variable and sampling it (together with the other parameters) from the posterior distribution with RJMCMC. We basically apply the same approach. The difference is that our emission probabilities are not distributions over individual nucleotides, but over clusters of tree topologies; the latter have been obtained in the pre-processing step described in Section 2.2. The posterior distributions estimated with the proposed inference scheme do not only indicate the maximum a posteriori number of different homogeneous regions and their locations, but also indicate the uncertainty in the prediction.

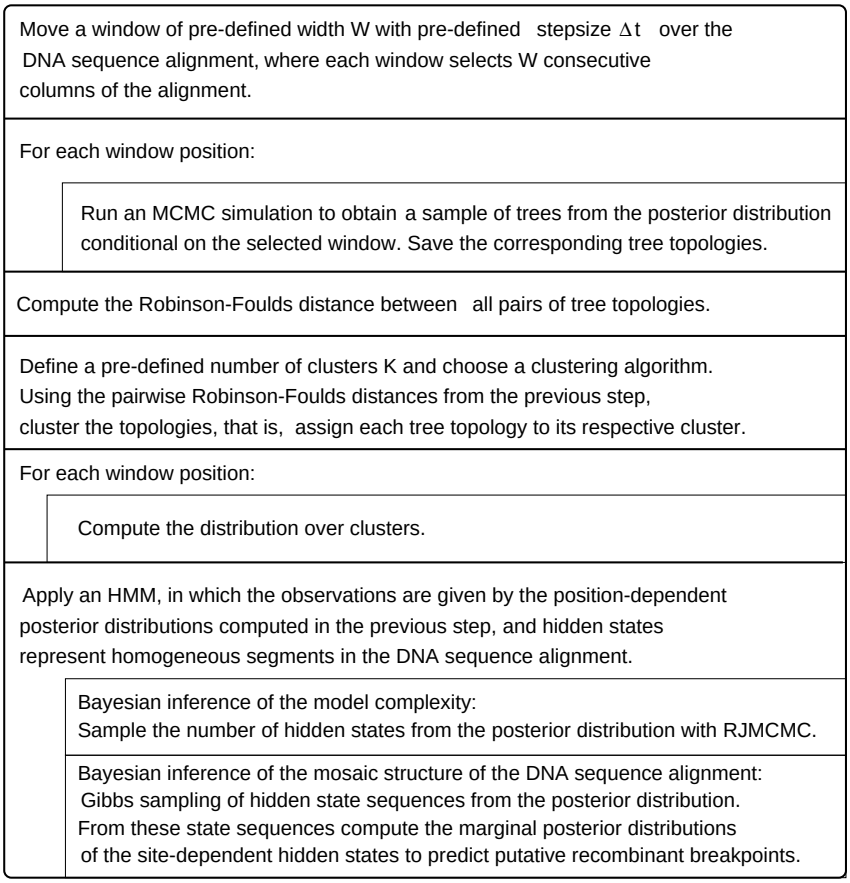


Figure 2.9: **Pseudo code of the combination with HMMs.** The figure summarizes the various steps involved in combining the pruned PDM measure with Bayesian HMMs.

Findings

We discussed the results obtained on various synthetic and real-world DNA sequence alignments in [DH17]. On synthetic DNA sequence alignment, the method correctly predicted the number as well as the location of the recombinant regions, and the predictions were robust with respect to changes in the prior distribution and the pre-processing scheme. In addition, the method succeeded in clearly distinguishing between recombination and rate variation[3]. On real DNA sequence alignments from maize actin genes, Hepatitis-B virus and HIV-1 the results showed some variation with respect to changes in the prior distribution and the pre-processing parameters. However, the results reported in the literature show the same variation, in response to employing different detection methods, or using different parameters of the same detection method. The differences between the predicted changepoints in our study were rather minor, and for each set of predicted changepoints we could find some method in the literature that supported these predictions. In summary, the results of our simulation studies suggested that the inference scheme either identified the correct number

[3]To be discussed in Chapter 4.

and location of the recombinant regions, or made predictions that were consistent with the results reported in the literature. This suggests that the proposed method is relatively robust and offers a viable exploration scheme for finding evidence of recombination in DNA sequence alignments.

2.4 Discussion

Although the individual simulations for different windows are based on rigorous Bayesian phylogenetic modelling, the overall scheme depends on various heuristic parameters: the window size, the number of clusters in the pruning scheme, and (for the combination with Bayesian HMMs) the prior distribution on the number of segments. Hence, window-based methods provide a tool for exploratory data analysis and hypothesis generation, rather than validation. By varying the prior distribution and the pre-processing parameters, we can control the resolution and the degree of conservatism of the proposed detection method. The time-consuming step appears to be the repeated run of MCMC simulations on different subsets of the alignment (windows). However, these simulations can be run in parallel, and the methods are therefore ideally suited for modern high-performance computer clusters. This substantially reduces the computing times over the more accurate models that will be discussed in the next two chapters.

Chapter 3

Detecting Recombination with Phylogenetic Hidden Markov Models

This chapter summarizes my work in [DH15,DH25,DH26,DH28] and describes an approach based on phylogenetic HMMs. The method draws on the observation that interspecific recombination usually leads to a change of the underlying phylogenetic tree topology. The idea is to introduce a hidden state that represents the tree topology at a given site. A state transition from one topology into another corresponds to a recombination event. To introduce correlations between adjacent sites, the hidden states are given a Markovian dependence structure. Thus, the standard model of a phylogenetic tree is generalized by the combination of two probabilistic models: a taxon graph (phylogenetic tree) representing the relationships among the taxa, and a site graph (HMM) representing dependencies between different sites in the DNA sequence alignments. Changepoints of mosaic segments in the alignment are predicted by state transitions in the site graph. While this method can only deal with a small number of sequences simultaneously, it has the potential to predict the locations and changepoints of recombinant regions more accurately than what can be achieved with most existing techniques.

3.1 Introduction

Phylogenetic hidden Markov models (HMMs) are motivated by the older approach of RECPARS [63], which is based on maximum parsimony (see Section 1.3) and was briefly mentioned in Sec-

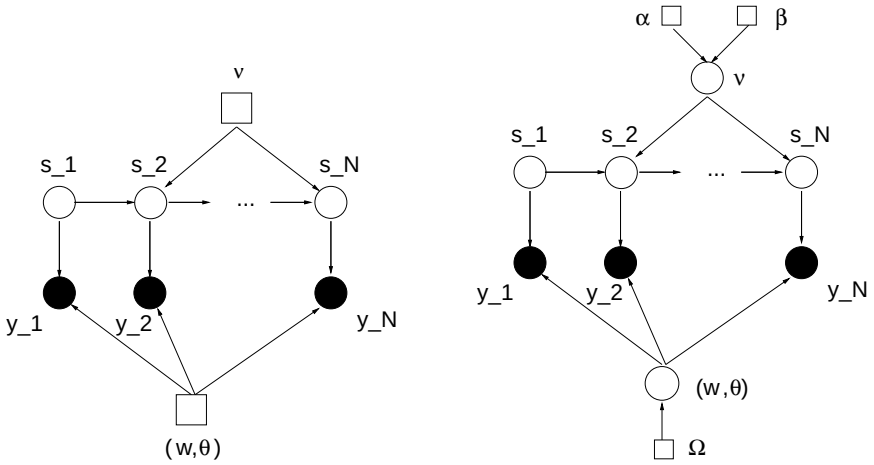


Figure 3.1: **Modelling recombination with hidden Markov models.** Positions in the model, labelled by the subscript t, correspond to sites in the DNA sequence alignment. Black nodes represent observed random variables; these are the columns in the DNA sequence alignment. White nodes represent hidden states; these are the different tree topologies, shown (for four sequences) in Figure 3.2. Arrows represent conditional dependencies. Squares represent parameters of the model. The probability of observing a column vector $\mathbf{y}_t$ at position t in the DNA sequence alignment depends on the tree topology S_t, the vector of branch lengths $\mathbf{w}$, and the parameters of the nucleotide substitution model $\boldsymbol{\theta}$. The tree topology at position t depends on the topologies at the adjacent sites, S_{t-1} and S_{t+1}, and the recombination parameter ν. *Left:* In the maximum likelihood approach, ν, $\mathbf{w}$, and $\boldsymbol{\theta}$ are parameters that have to be estimated. *Right:* In the Bayesian approach, ν, $\mathbf{w}$, and $\boldsymbol{\theta}$ are random variables. The prior distribution for ν is a beta distribution with hyperparameters α and β. The prior distributions for the remaining parameters depend on some hyperparameters, generically denoted by Ω. The parameters ν, $\mathbf{w}$, and $\boldsymbol{\theta}$ are sampled from the posterior distribution with Markov chain Monte Carlo (MCMC).

tion 1.7. Phylogenetic HMMs are based on the probabilistic framework described in Section 1.4 and thereby overcome the two fundamental shortcomings of RECPARS. Firstly, they avoid the problem of inconsistency that is inherent in maximum parsimony; see Section 1.3. Secondly, as opposed to methods based on maximum parsimony, they allow all parameters to be consistently inferred from the data. I have provided a detailed comparison between these two approaches in [DH28].

The essential concept of phylogenetic HMMs is illustrated in Figures 3.1-3.2. The left panel of Figure 3.1 depicts the structure of the model as a probabilistic graphical model. White nodes represent hidden states, S_t, which have direct interactions only with the states at adjacent sites, S_{t-1} and S_{t+1}. Black nodes represent columns in the DNA sequence alignment, $\mathbf{y}_t$. The squares represent the model parameters, which are the branch lengths of the phylogenetic trees, $\mathbf{w}$, the parameters of the nucleotide substitution model, $\boldsymbol{\theta}$, and the recombination parameter, ν; see Figures 3.2 and 3.3. The joint probability of the DNA sequence alignment, $\mathcal{D} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, and the sequence of hidden states,[1] $\mathbf{S} = (S_1, \ldots, S_N)$, factorizes, by the standard expansion rule for Bayesian networks

[1]Note the difference between $\mathbf{S}$ (a state sequence) and S or S_t (an individual hidden state), as explained in the previous footnote.

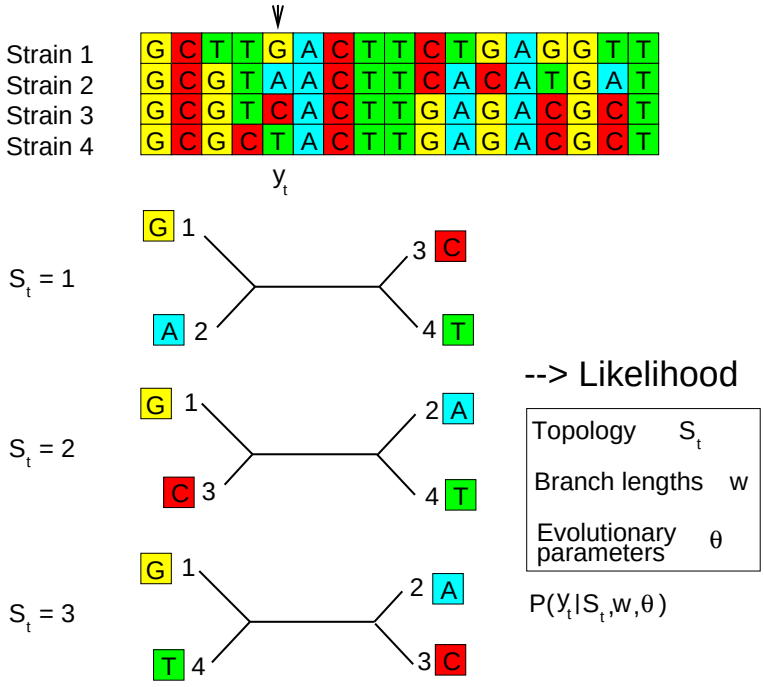


Figure 3.2: **Probabilistic approach to phylogenetics and modelling recombination.** For a given column $\mathbf{y}_t$ in the sequence alignment, a probability $P(\mathbf{y}_t|S_t, \mathbf{w}, \boldsymbol{\theta})$ can be computed, which depends on the tree topology S_t, the vector of branch lengths $\mathbf{w}$, and the parameters of the nucleotide substitution model $\boldsymbol{\theta}$. In the presence of recombination, the tree topology can change and thus becomes a random variable that depends on the site label t. For four taxa, there are three different tree topologies. The vectors $\mathbf{w}$ and $\boldsymbol{\theta}$ are accumulated vectors, as defined in the paragraph above equation (3.3).

(see e.g. Chapter 2 in [DH1]):

$$\begin{aligned} P(\mathcal{D}, \mathbf{S}|\mathbf{w}, \boldsymbol{\theta}, \nu) &= P(\mathbf{y}_1, \ldots, \mathbf{y}_N, S_1, \ldots, S_N|\mathbf{w}, \boldsymbol{\theta}, \nu) \\ &= P(S_1) \prod_{t=1}^{N} P(\mathbf{y}_t|S_t, \mathbf{w}, \boldsymbol{\theta}) \prod_{t=2}^{N} P(S_t|S_{t-1}, \nu) \end{aligned} \tag{3.1}$$

$P(S_1)$ is the marginal distribution over the hidden states, that is, the K possible tree topologies. The $P(S_t|S_{t-1}, \nu)$ are the transition probabilities, which depend on the recombination parameter $\nu \in [0, 1]$, as illustrated in Figure 3.3. The functional form chosen in [DH25,DH26] is given by

$$P(S_t|S_{t-1}, \nu_S) = \nu^{\delta(S_t, S_{t-1})} \left(\frac{1-\nu}{K-1}\right)^{[1-\delta(S_t, S_{t-1})]} \tag{3.2}$$

where $\delta(S_t, S_{t-1})$ denotes the Kronecker delta symbol. The parameter ν defines the probability that on moving from a site in the DNA sequence alignment to an adjacent site, no topology change occurs. If a topology change occurs – corresponding to a recombination event – all transitions into other topologies are assumed to be equally likely. The $P(\mathbf{y}_t|S_t, \mathbf{w}, \boldsymbol{\theta})$ are the emission probabilities,

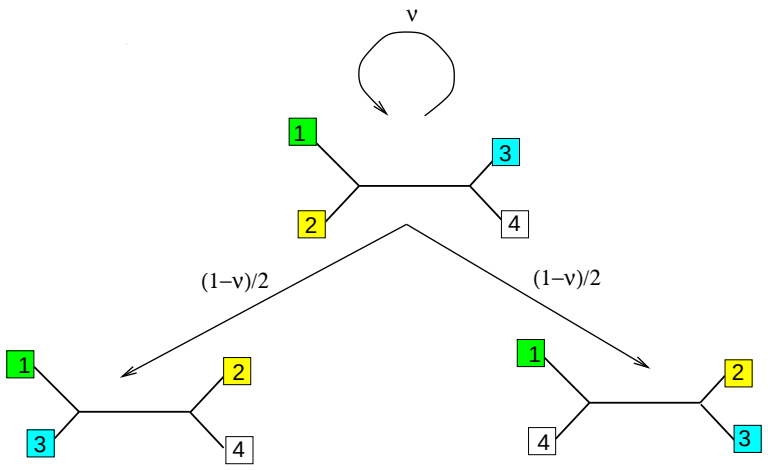


Figure 3.3: **Transition probabilities.** The hidden states of the HMM represent different tree topologies, and state transitions correspond to recombination events. The transition probability ν is the probability that on moving from a site in the DNA sequence alignment to an adjacent site, no topology change occurs. If a topology change occurs, we assume that, *a priori*, all transitions are equally likely.

which depend on the vector of branch lengths, $\mathbf{w}$, and the parameters of the nucleotide substitution model, $\boldsymbol{\theta}$. For example, for the Kimura model illustrated in Figures 1.10 and 1.12 , $\boldsymbol{\theta}$ is given by the transition-transversion ratio. For the HKY85 model, mentioned in Section 1.4, , $\boldsymbol{\theta}$ is a vector containing the transition-transversion ratio and the equilibrium probabilities of the nucleotides.[2] The computation of $P(\mathbf{y}_t|S_t, \mathbf{w}, \boldsymbol{\theta})$ was discussed in Section 1.4. Note that while the standard likelihood approach to phylogenetics, discussed in Chapter 1, assumes that one tree topology S applies to the whole sequence alignment, equation (3.1) allows for topology changes, corresponding to recombination events. The best prediction of the recombinant regions and their changepoints is given by

$$\begin{aligned} \hat{\mathcal{S}} &= \text{argmax}_{\mathbf{S}} P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \boldsymbol{\theta}, \nu) \\ &= \text{argmax}_{S_1,\ldots,S_N} P(S_1, \ldots, S_N|\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{w}, \boldsymbol{\theta}, \nu) \end{aligned} \tag{3.3}$$

which is computed with the Viterbi algorithm for HMMs [109]. The marginal probability $P(S_1)$ and the parameters $\mathbf{w}$, $\boldsymbol{\theta}$, and ν need to be estimated.

[2] In fact, both $\mathbf{w}$ and $\boldsymbol{\theta}$ are dependent on the hidden state, so more precisely, the emission probabilities should be written as $P(\mathbf{y}_t|S_t, \mathbf{w}_{S_t}, \boldsymbol{\theta}_{S_t})$. To avoid clutter in the notation, I prefer to use the accumulated vectors $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_K)$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ and define: $P(\mathbf{y}_t|S_t, \mathbf{w}_{S_t}, \boldsymbol{\theta}_{S_t}) = P(\mathbf{y}_t|S_t, \mathbf{w}, \boldsymbol{\theta})$. This means that S_t indicates which subvectors of $\mathbf{w}$ and $\boldsymbol{\theta}$ apply.

3.2 Maximum Likelihood

The application of phylogenetic HMMs to the detection of recombination was not my idea; in fact, the approach was first proposed in [95]. However, the authors did not find a satisfactory solution to the inference problem. In fact, they described the problem as a "chicken and egg" paradox, as illustrated in Figure 3.4. In order to infer the parameters of the recombinant and dominant phylogenetic trees, most notably their branch lengths $\mathbf{w}$, one needs to know where the recombinant regions are. However, it is the absence of this knowledge that motivates the application of the phylogenetic HMM in the first place. Inferring the branch lengths for each tree from the whole DNA sequence alignment is methodologically inconsistent, as illustrated in the left panel of Figure 3.4: the estimation of the branch lengths of the recombinant tree, corresponding to the centre segment of the sequence alignment, would be dominated by the flanking non-recombinant regions, which should actually be excluded. Conversely, the estimation of the branch lengths of the dominant tree, corresponding to the flanking regions of the alignment, would be distorted by the presence of the recombinant region. The authors addressed this issue by incorporating into the estimation procedure a window-based technique of the type discussed in the previous chapter, but this approach is heuristic and suboptimal. My contribution in [DH28] was to demonstrate that all parameters can be inferred consistently in a maximum likelihood sense with the expectation maximization (EM) algorithm first proposed in [33]. A complete description of this approach can be found in [DH28]. In a nutshell, the algorithm works as follows. In the maximization step (M-step), the branch lengths and nucleotide substitution parameters are optimized by maximizing

$$A(\mathbf{w}, \boldsymbol{\theta}) \;=\; \sum_{t=1}^{N} \sum_{S_t} Q(S_t) \log P(\mathbf{y}_t | S_t, \mathbf{w}, \boldsymbol{\theta}) \tag{3.4}$$

where $Q(S_t)$ is a site-dependent probability distribution over the states S_t to be discussed shortly. The optimization problem is NP-hard. However, standard greedy search routines used in likelihood-based phylogenetic inference can be applied here. In fact, any maximum likelihood phylogenetic inference package can be used, like DNAML in PHYLIP [42].[3] The only modification required is a state-dependent weighting of the sites in the DNA sequence alignment with the factor $Q(S_t)$, as illustrated in Figure 3.4. The factors $Q(S_t)$ are estimated in the expectation step (E-step) of the algorithm, using the forward–backward algorithm for HMMs [109]. The cycles of E and M steps have to be iterated and are guaranteed to converge to a stationary (i.e. zero gradient) configuration of the likelihood. This is usually a local maximum, and the algorithm is hence run repeatedly, starting from different initializations. There are analogous M-steps for the marginal probabilities $P(S_t)$ and the recombination parameter ν, and the complete mathematical description of the algorithm can be found in [DH28].

The maximum likelihood approach has the disadvantage that the optimal sequence of hidden states, $\hat{\mathcal{S}}$, given by (3.3), depends on the parameters $\mathbf{w}$, $\boldsymbol{\theta}$. These parameters were estimated from the

[3] Available from `http://evolution.genetics.washington.edu/phylip/`.

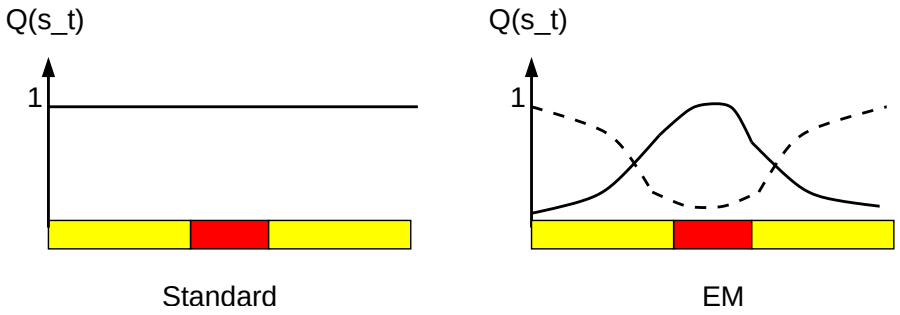


Figure 3.4: **Nucleotide weighting schemes.** The bottom of each figure represents a multiple DNA sequence alignment with a recombinant zone in the central segment. The *left panel* shows the standard method of phylogenetic inference. The tree parameters are estimated from the whole sequence alignment, which corresponds to a uniform weight of 1 for all sites. The *right panel* illustrates the EM algorithm. The solid line shows the site-dependent weights $Q(S_t = T_R)$ for the recombinant topology T_R, the dashed line represents the weights for the non-recombinant topology T_0 : $Q(S_t = T_0)$. Note that the weights $Q(S_t)$ are updated automatically in every iteration of the optimization procedure (in the E-step of the EM algorithm).

sequence alignment, which renders the approach susceptible to overfitting. Within the frequentist paradigm of inference, we have to test the null hypothesis of no recombination against the alternative hypothesis that a recombination event has occurred. For the practical implementation, we can use nonparametric or parametric bootstrapping, as described in [DH28]. These approaches are computationally expensive, though, because they require us to repeat the iterative optimization procedure of the EM algorithm on many (typically a hundred) bootstrap replicas of the sequence alignment; see Figure 1.15. An alternative would be to resort to asymptotic model selection scores, like the likelihood ratio test [66] or the BIC score [128]. However, for complex models and comparatively small data sets, these scores cannot be assumed to be reliable, and their use is not recommended. Motivated by Figure 1.15, I have therefore switched to a Bayesian approach, as described in the following section.

3.3 Bayesian Approach

Within the Bayesian framework, motivated in Section 1.5, the prediction of the optimal state sequence $\mathbf{S}$ should be based on the posterior distribution $P(\mathbf{S}|\mathcal{D})$, which requires the remaining parameters to be integrated out:

$$P(\mathbf{S}|\mathcal{D}) = \int P(\mathbf{S}, \mathbf{w}, \boldsymbol{\theta}, \nu|\mathcal{D}) d\mathbf{w} d\boldsymbol{\theta} d\nu \tag{3.5}$$

In principle this avoids the overfitting scenario mentioned above and removes the need for a separate hypothesis test. The difficulty, however, is that the integral in (3.5) is analytically intractable and needs to be numerically approximated with Markov chain Monte Carlo (MCMC). A complete

description of this scheme can be found in [DH25,DH26]. Below, I shall briefly discuss the choice of prior distribution and the implementation of the MCMC method.

Prior distributions

Inherent in the Bayesian framework is the choice of prior distributions for all model parameters, as illustrated in Figure 3.1, right. The approach in [DH25,DH26] makes the common assumption of parameter independence [60], $P(\nu, \mathbf{w}, \boldsymbol{\theta}) = P(\nu)P(\mathbf{w})P(\boldsymbol{\theta})$. Thus, with (3.1), we obtain for the joint distribution:

$$P(\mathcal{D}, \mathbf{S}, \mathbf{w}, \boldsymbol{\theta}, \nu) \;=\; P(S_1)P(\mathbf{w})P(\boldsymbol{\theta})P(\nu)\prod_{t=1}^{N} P(\mathbf{y}_t|S_t, \mathbf{w}, \boldsymbol{\theta})\prod_{t=2}^{N} P(S_t|S_{t-1}, \nu) \tag{3.6}$$

Due to the absence of specific biological knowledge about the nature of recombination processes, the prior distributions are chosen rather vague. Also, prior distributions are chosen either conjugate, where possible, or uniform, but proper (that is, restricted to a finite interval). The conjugate prior for the recombination parameter ν is a beta distribution

$$\mathcal{B}(\nu) \;=\; \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\nu^{\alpha-1}(1-\nu)^{\beta-1} \tag{3.7}$$

whose shape is determined by two hyperparameters α and β. For $\alpha = \beta = 1$, the distribution in (3.7) is the uniform distribution over the unit interval. It is straightforward to set α and β so as to incorporate prior knowledge about the global frequency of recombination events. *A priori*, the branch lengths $\mathbf{w}$ are assumed to be uniformly distributed in the interval $[0, 1]$.[4] In [DH25,DH26], the prior on S_1 was chosen uniform. A more general approach is to treat $P(S_1)$ itself as a parameter vector and assign it a conjugate Dirichlet prior. Finally, the prior on $\boldsymbol{\theta}$ depends on the nucleotide substitution model and is discussed in [DH25,DH26].

The MCMC scheme

Ultimately, we are interested in the marginal posterior distribution of the state S_t at a given site t in the alignment:

$$P(S_t|\mathcal{D}) \;=\; \sum_{S_1}\cdots\sum_{S_{t-1}}\sum_{S_{t+1}}\cdots\sum_{S_N}\int P(\mathbf{S}, \mathbf{w}, \boldsymbol{\theta}, \nu|\mathcal{D})d\mathbf{w}d\boldsymbol{\theta}d\nu \tag{3.8}$$

This requires a marginalization over the model parameters and the states at the other sites. The numerical approximation is to sample from the joint posterior distribution

$$P(\mathbf{S}, \mathbf{w}, \boldsymbol{\theta}, \nu|\mathcal{D}) \tag{3.9}$$

[4] Fixing an upper bound on the branch lengths is necessary to avoid the use of an improper prior, for which the MCMC scheme might not converge. Since for real DNA sequence alignments branch lengths are unlikely to approach values as large as 1, this restriction should not cause any difficulties.

and then to proceed as discussed in Section 2.1 and discard the model parameters and the states at the other sites. Sampling from the joint posterior distribution follows a Gibbs sampling scheme [25]. Applied to (3.9), this method samples each parameter group separately conditional on the others. So if the superscript (i) denotes the ith sample of the Markov chain, we obtain the $(i+1)$th sample as follows:

$$\begin{aligned}
\mathbf{S}^{(i+1)} &\sim P(\cdot|\mathbf{w}^{(i)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}, \mathcal{D}) && (3.10)\\
\mathbf{w}^{(i+1)} &\sim P(\cdot|\mathbf{S}^{(i+1)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}, \mathcal{D}) && (3.11)\\
\boldsymbol{\theta}^{(i+1)} &\sim P(\cdot|\mathbf{S}^{(i+1)}, \mathbf{w}^{(i+1)}, \nu^{(i)}, \mathcal{D}) && (3.12)\\
\nu^{(i+1)} &\sim P(\cdot|\mathbf{S}^{(i+1)}, \mathbf{w}^{(i+1)}, \boldsymbol{\theta}^{(i+1)}, \mathcal{D}) && (3.13)
\end{aligned}$$

The order of these sampling steps, which will be discussed in the remainder of this subsection, is arbitrary.

The posterior distribution of the recombination parameter ν can be shown to depend on the state sequences S via the sufficient statistics $\Psi = \sum_{t=1}^{N-1} \delta_{S_t,S_{t+1}}$ and is given by

$$P(\nu|\mathcal{D}, \mathbf{S}, \mathbf{w}, \boldsymbol{\theta}) \;=\; \mathcal{B}(\nu|\Psi + \alpha, N - \Psi + \beta - 1) \tag{3.14}$$

where $\mathcal{B}$ is the beta distribution (3.7), from which sampling is straightforward [121].

For sampling the state sequences $\mathbf{S}$, the approach pursued in [DH25,DH26] is a Gibbs-within-Gibbs scheme, where each state S_t is sampled separately conditional on the others. Conditional independence relations in HMMs are exploited to simplify the computations:

$$\begin{aligned}
P(S_t|S_1, \ldots, S_{t-1}, S_{t+1}, \ldots, S_N, \mathcal{D}, \mathbf{w}, \boldsymbol{\theta}, \nu) &= P(S_t|S_{t-1}, S_{t+1}, \mathbf{y}_t, \mathbf{w}, \boldsymbol{\theta}, \nu) && (3.15)\\
&\propto P(S_{t+1}|S_t, \nu) P(S_t|S_{t-1}, \nu) P(\mathbf{y}_t|S_t, \mathbf{w}, \boldsymbol{\theta})
\end{aligned}$$

where $P(S_t|S_{t-1}, \nu)$ and $P(S_{t+1}|S_t, \nu)$ are given by (3.2). Note that the last expression in (3.15) is easily normalized to give a proper probability distribution, from which sampling is straightforward (since $S_t \in \{1, \ldots, K\}$ is discrete). This approach of Gibbs-within-Gibbs sampling was also used in [114]. However, I later learned that mixing and convergence of the Markov chains can be substantially improved by adopting a dynamic programming scheme, which I call the stochastic forward-backward algorithm. I relegate the description of this scheme to the appendix at the end of this chapter (Section 3.6). The stochastic forward-backward algorithm is effectively a combination of the forward algorithm in HMMs [109] with a backward sampling procedure. It can also be regarded as a modified Viterbi algorithm [109], in which the backtracking optimization routine is replaced by a backtracking sampling routine. As opposed to [114], Boys et al. [19] had conjectured that its mixing and convergence properties should be superior to the Gibbs-within-Gibbs scheme of (3.15), which we empirically confirmed in [DH52].

Finally, the remaining parameters, $\mathbf{w}$ and $\boldsymbol{\theta}$, are sampled with the Metropolis–Hastings algorithm, as described in [DH25,DH26]. The overall procedure is thus of the form of a Metropolis–Hastings–

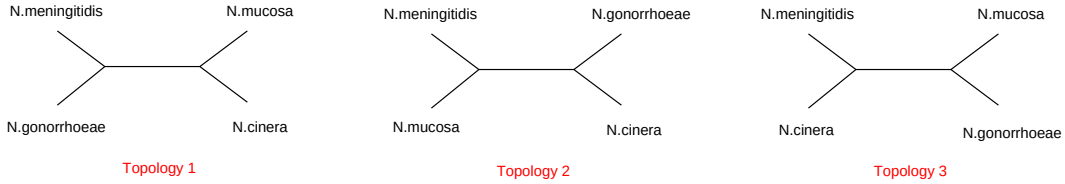


Figure 3.5: **Different phylogenetic tree topologies for four strains of *Neisseria*.** Zhou and Spratt [157] investigated the 787-nucleotide *argF* gene sequence alignment of four *Neisseria* strains: (1) *N.gonorrhoeae* (X64860), (2) *N.meningitidis* (X64866), (3) *N.cinerea* (X64869), and (4) *N.mucosa* (X64873); GenBank/EMBL accession numbers are given in brackets. The figure shows the three possible phylogenetic tree topologies.

within–Gibbs scheme. The algorithm has been implemented in a publicly available software package [DH11,DH22].

3.4 Findings

From the many applications and comparative evaluations presented in [DH25,DH26,DH28], I pick a particular interesting one for illustration purposes. One of the first indications for interspecific recombination was found in the bacterial genus *Neisseria* [92]. Zhou and Spratt [157] investigated the 787-nucleotide *argF* gene sequence alignment of four strains, shown in Figure 3.5. In the main part of the alignment, *N.meningitidis* is grouped with *N.gonorrhoeae*, corresponding to topology $S_t = 1$ in Figure 3.5. However, Zhou and Spratt [157] found two anomalous regions in the DNA alignment. Between positions $t = 1$ and 202, they found a phylogenetic tree with a different topology, corresponding to topology $S_t = 3$ in Figure 3.5. They also found a more diverged region with the same topology, topology $S_t = 1$ in Figure 3.5, between positions $t = 507-$ and 538. The situation is illustrated in Figure 3.6.

The bottom panels of Figure 3.6 show the prediction of $P(S_t|\mathcal{D})$, where the subfigure in the bottom left was obtained with the maximum likelihood HMM method of Section 3.2, henceforth called HMM-ML, and the subfigure in the bottom right with the Bayesian HMM method of Section 3.3, henceforth referred to as HMM-Bayes. Both methods agree in predicting a sharp transition from topology $S_t = 3$ to $S_t = 1$ at changepoint $t = 202$, which is in agreement with the findings in [157]. Both methods also agree in predicting a short recombinant region of the same topology change at the end of the alignment. This region was not reported in [157]. Since it is very short – only about 20 nucleotides long – it may not be detectable by other, less accurate methods.

Differences between the predictions of HMM-ML and HMM-Bayes are found in the middle of the alignment, where two further changepoints occur at sites $t = 506$ and $t = 537$. This is in agreement with [157]. However, while Zhou and Spratt [157] suggested that the region between $t = 506$ and

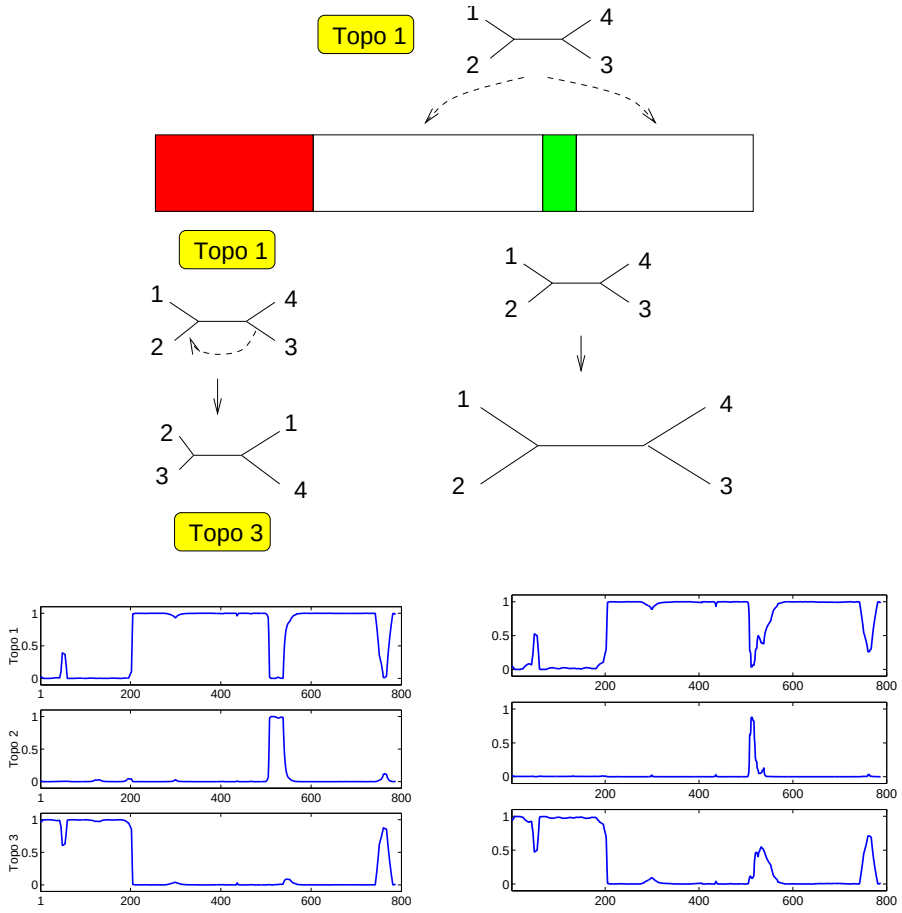


Figure 3.6: **Recombination in Neisseria.** *Top:* According to Zhou and Spratt [157], a recombination event corresponding to a transition from topology 1 to topology 3, as defined in Figure 3.5, has affected the first 202 nucleotides of the DNA sequence alignment of four strains of *Neisseria*. A second more diverged segment seems to be the result of rate variation. *Bottom:* Prediction with the phylogenetic HMM. The figure contains two subfigures. Each subfigure is composed of three graphs, which plot the marginal posterior probabilities of the three topologies defined in Figure 3.5 against the position in the DNA sequence alignment. *Left panel:* Prediction with the maximum likelihood method of Section 3.2. *Right panel:* Prediction with Bayesian learning using the MCMC scheme of Section 3.3.

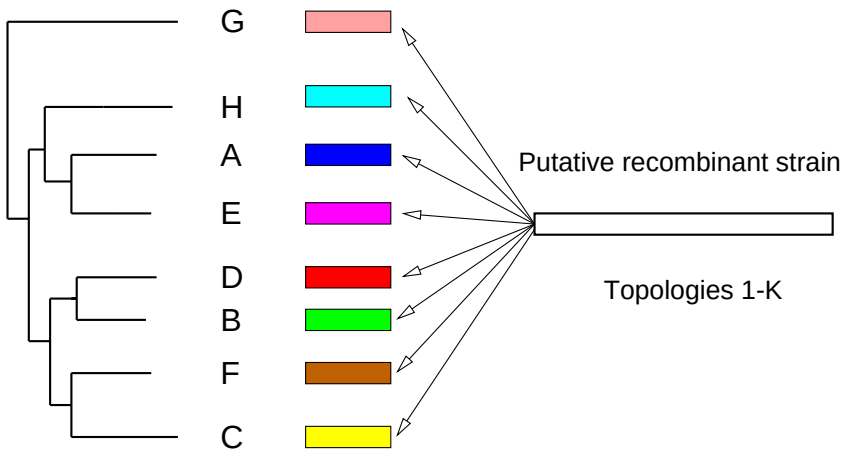


Figure 3.7: **Computational complexity of the phylogenetic HMM.** Given a set of non-recombinant strains with a known phylogenetic tree, different segments of the DNA sequence of a putative recombinant strain will be grouped with different leaves of the tree. This leads to n candidate tree topologies to be included as hidden states in the phylogenetic HMM. Note that this number is considerably less than the total number of unrooted tree topologies for $(n+1)$ taxa, which is $(2n-3)!!$.

$t = 537$ seems to be the result of rate heterogeneity, HMM-ML predicts a recombination event with a clear transition from topology $S_t = 1$ into $S_t = 2$. This seems to be the result of overfitting: since the distribution of the nucleotide column vectors $\mathbf{y}_t$ in the indicated region is significantly different from the rest of the alignment, modelling this region with a different hidden state can increase the likelihood although the hidden state itself (topology $S_t = 2$) might be ill-matched to the data. This deficiency is redeemed with HMM-Bayes, whose prediction is shown in Figure 3.6, bottom right. The critical region between sites $t = 506$ and $t = 537$ is again identified, indicated by a strong drop in the posterior probability for the predominant topology, $P(S_t = 1|\mathcal{D})$. However, the uncertainty in the nature of this region is indicated by a distributed representation, where both alternative hidden states, $S_t = 2$ and $S_t = 3$, are assigned a significant probability mass. With the prediction of this uncertainty, HMM-Bayes also indicates a certain model misspecification inherent in the current scheme – the absence of hidden states for representing different nucleotide substitution rates. I will discuss a generalization of the phylogenetic HMM to address this shortcoming in the next chapter.

3.5 Computational Complexity

The phylogenetic HMM assumes that each potentially possible phylogenetic tree topology constitutes a separate hidden state of the HMM. As mentioned in the Introduction chapter, there are $(2n-5)!!$ different unrooted tree topologies for n taxa. Hence, the number of hidden states in the HMM increases superexponentially with the number of sequences in the alignment. One approach is to restrict the phylogenetic HMM to small sets of taxa and use it as a fine resolution method in conjunction with the window-based approaches that were discussed in the previous chapter; the latter are used for coarse-grain screening so as to identify sets of putative recombinant strains. I have applied this procedure in [DH15]. The other approach is illustrated in Figure 3.7. If we have a set of non-recombinant strains, then different sequence segments of a putative recombinant strain will be grouped with different leave nodes of the phylogenetic tree formed by the non-recombinant strains, and the number of potentially possible tree topologies reduces from $(2n-3)!!$ to n. I have investigated this approach with Alexander Mantzaris, who has demonstrated its feasibility in his MSc dissertation [91].

3.6 Appendix

Hidden state sequences can be sampled from the conditional distribution in (3.10) with the following dynamic programming scheme (which I called the stochastic forward-backward algorithm in [DH52]). Define

$$\alpha_t(S_t) \;=\; P(\mathbf{y}_1, \ldots, \mathbf{y}_t, S_t) \tag{3.16}$$

which is the function computed in the forward pass of the forward-backward algorithm for HMMs; see, for instance, [109]. Now,

$$\begin{aligned}
& P(S_t | S_{t+1}, \ldots, S_N, \mathbf{y}_1, \ldots, \mathbf{y}_N) \\
\propto\; & P(S_t, S_{t+1}, \ldots, S_N, \mathbf{y}_1, \ldots, \mathbf{y}_N) \\
=\; & P(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N, S_{t+1}, \ldots, S_N | S_t, \mathbf{y}_1, \ldots, \mathbf{y}_t) P(S_t, \mathbf{y}_1, \ldots, \mathbf{y}_t) \\
=\; & P(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N, S_{t+1}, \ldots, S_N | S_t) \alpha_t(S_t) \\
=\; & P(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N, S_{t+2}, \ldots, S_N | S_{t+1}) P(S_{t+1} | S_t) \alpha_t(S_t) \\
\propto\; & P(S_{t+1} | S_t) \alpha_t(S_t)
\end{aligned} \tag{3.17}$$

The simplifications carried out here follow directly from the independence relations in HMMs. The last step follows from the fact that the first term in the second-last line is independent of S_t and therefore cancels out in the normalization:

$$P(S_t = \tau_k | S_{t+1}, \ldots, S_N, \mathbf{y}_1, \ldots, \mathbf{y}_N) \;=\; \frac{P(S_{t+1} | S_t = \tau_k) \alpha_t(S_t = \tau_k)}{\sum_i P(S_{t+1} | S_t = \tau_i) \alpha_t(S_t = \tau_i)} \tag{3.18}$$

Obviously, any scaling constant also cancels out in the normalization; hence replacing $\alpha_t(S_t)$ by some scaled version for numerical stabilization of the forward algorithm will not affect the result. The algorithm is initialized by drawing the final state, S_N, from the following distribution:

$$P(S_N = \tau_k | \mathbf{y}_1, \ldots, \mathbf{y}_N) \;=\; \frac{\alpha_N(S_N = \tau_k)}{\sum_i \alpha_N(S_N = \tau_i)} \tag{3.19}$$

The overall algorithm can thus be summarized as follows: First, run the (scaled) forward algorithm [109]. Next, sample S_N from (3.19). Finally, sample the remaining states $S_{N-1}, \ldots, S_1$ recursively from (3.18). Note that at the end of this recursion, a whole state sequence $\mathbf{S} = (S_1, \ldots, S_N)$ has been sampled from $P(\mathbf{S}|\mathbf{w}^{(i)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}, \mathcal{D})$ of equation (3.10). In [DH52] we showed that as a consequence of the improved mixing and convergence, the computational costs could be reduced by an order of magnitude over the Gibbs-within-Gibbs scheme of (3.15).

Chapter 4

Distinguishing between Rate Heterogeneity and Recombination

A shortcoming of phylogenetic HMMs is their inability to differentiate between recombination and rate heterogeneity. I address this issue in the present chapter, which summarizes my work in [DH9,DH10,DH20,DH46]. The idea is to extend the HMM approach to a factorial HMM (FHMM). The states of the first hidden chain represent tree topologies, as before, and transitions between these states are indicative of recombination events. The states of the second independent hidden chain represent different global scaling factors of the branch lengths, and transitions between these rate states indicate rate heterogeneity, potentially related to variations in the selective pressure. I discuss different Bayesian inference schemes based on Gibbs sampling [DH20] and transdimensional reversible jump MCMC [DH9], Bayesian model selection [DH10] and within-codon effects [DH46].

4.1 Introduction and Method

The results illustrated in Figure 3.6 have revealed a fundamental limitation intrinsic to the phylogenetic HMM. While it tends to detect the location of recombinant regions and their demarcation points more accurately than most other techniques, it inherently fails to distinguish between recombination and rate heterogeneity. Hence, genomic regions under different selective pressure tend to be erroneously predicted as resulting from recombination events. The nature of rate heterogeneity is illustrated in Figure 4.1, and this chapter will discuss how to extend the model so as to detect it

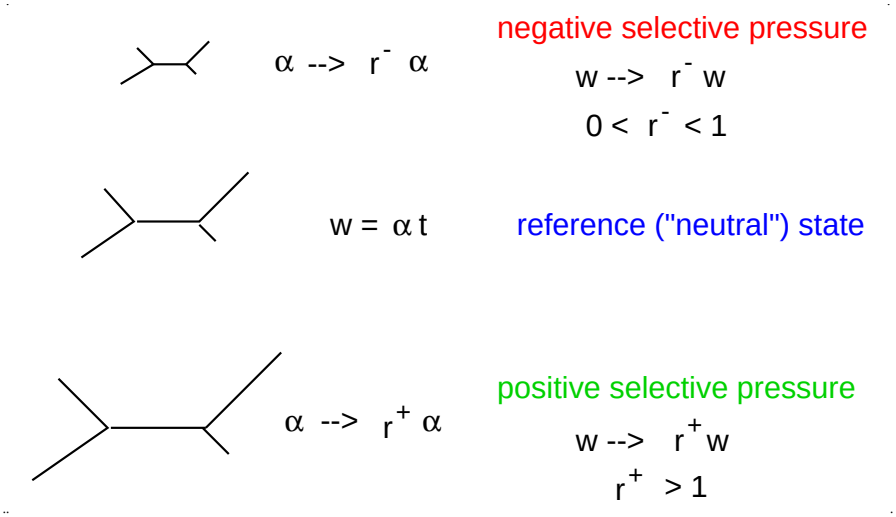


Figure 4.1: **Simple illustration of rate heterogeneity.** As discussed in the Introduction chapter, the branch lengths of the phylogenetic tree are given by the product of a nucleotide substitution rate α and physical time t, as shown in equation (1.1). Under negative selective pressure, the mutation rate is reduced, and the branch lengths shrink by a factor $0 < r < 1$. Under positive selective pressure, the mutation rate is increased, and the branch lengths increase by a factor $r > 1$.

and distinguish it from recombination.

The basic idea is depicted in Figures 4.2 and 4.3. In addition to the hidden states that denote phylogenetic tree topologies, denoted by S_t, we introduce a separate *a priori* independent chain of hidden states, denoted by R_t, which denote different global scaling factors of the branch lengths. Transitions between S_t states are indicative of recombination. Transitions between R_t states indicate rate variation.

I shall briefly describe the mathematical model I proposed in [DH20] and [DH9], the latter in collaboration with Wolfgang Lehrach. The modelling starts by adopting an idea proposed in [139]. Different sites in the DNA sequence alignment are given separate branch length vectors $\mathbf{w}_t = (w_t^1, \ldots, w_t^{[2n-3]})$, for which prior independence between the individual branches w_t^i is assumed:

$$P(\mathbf{w}_t|\rho) = \prod_i P(w_t^i|\rho);\ P(w_t^i|\rho) = \rho^{-1}\exp(-w_t^i/\rho) \tag{4.1}$$

where ρ is a scaling factor determined by the rate state R_t. This prior is conjugate to the likelihood and makes integrating out the branch lengths analytically tractable:

$$P(\mathbf{y}_t|S_t,\rho) = \int P(\mathbf{y}_t|S_t,\mathbf{w}_{S_t})P(\mathbf{w}_t|\rho)d\mathbf{w}_t \tag{4.2}$$

See [139] for details. I shall discuss the nature of this approximation in more detail in Section 4.5, but for now it is just noted that as a consequence of this integration, the branch lengths of the

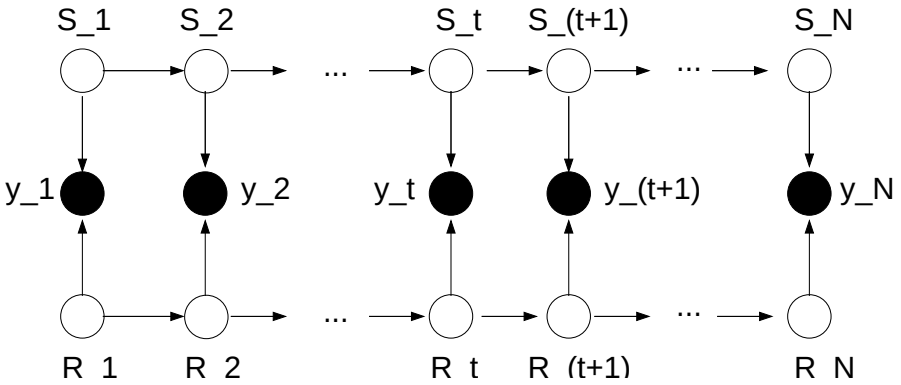


Figure 4.2: **Factorial hidden Markov model (FHMM).** Black nodes represent observed random variables; in the present applications these are columns of a DNA sequence alignment. Empty nodes represent latent variables or hidden states. There are two different types of hidden states, denoted by S_t and R_t. Both have a Markovian dependence structure and are *a priori* independent. In the present applications, the S_t states define the topology of a phylogenetic tree, while the R_t states are associated with different global scaling factors of the branches.

phylogenetic tree effectively disappear and the model simplifies. The complete likelihood factorizes as follows:

$$P(\mathcal{D}, \mathbf{S}, \mathbf{R}, \nu_S, \nu_R) = P(S_1)P(R_1)P(\nu_S)P(\nu_R) \prod_{t=1}^{N} P(\mathbf{y}_t|S_t, R_t) \prod_{t=2}^{N} P(S_t|S_{t-1}, \nu_S) \prod_{t=2}^{N} P(R_t|R_{t-1}, \nu_R) \tag{4.3}$$

where $P(\mathbf{y}_t|S_t, R_t)$ is given by equation (4.2), $\mathcal{D}$ denotes the DNA sequence alignment, which has N columns, $\mathbf{S}$ and $\mathbf{R}$ denote the sequences of topology states, S_t, and rate states, R_t, respectively, and the parameters ν_s and ν_r define the transition probabilities for topology and rate states, as in (3.2):

$$P(S_t|S_{t-1}, \nu_S) = \nu_S^{\delta(S_t, S_{t-1})} \left(\frac{1-\nu_S}{K-1}\right)^{[1-\delta(S_t, S_{t-1})]} \tag{4.4}$$

$$P(R_t|R_{t-1}, \nu_R) = \nu_R^{\delta(R_t, R_{t-1})} \left(\frac{1-\nu_R}{\tilde{K}-1}\right)^{[1-\delta(R_t, R_{t-1})]} \tag{4.5}$$

These parameters are given conjugate beta priors. Sampling from the joint posterior distribution follows a Gibbs sampling procedure [25], where each parameter group is iteratively sampled separately conditional on the others. So if the superscript (i) denotes the ith sample of the Markov chain, we obtain the $(i+1)$th sample as follows:

$$\mathbf{S}^{(i+1)} \sim P(\cdot|\mathbf{R}^{(i)}, \nu_S^{(i)}, \nu_R^{(i)}, \mathcal{D}) \tag{4.6}$$

$$\mathbf{R}^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)}, \nu_S^{(i)}, \nu_R^{(i)}, \mathcal{D}) \tag{4.7}$$

$$\nu_S^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)}, \mathbf{R}^{(i+1)}, \nu_R^{(i)}, \mathcal{D}) \tag{4.8}$$

$$\nu_R^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)}, \mathbf{R}^{(i+1)}, \nu_S^{(i+1)}, \mathcal{D}) \tag{4.9}$$

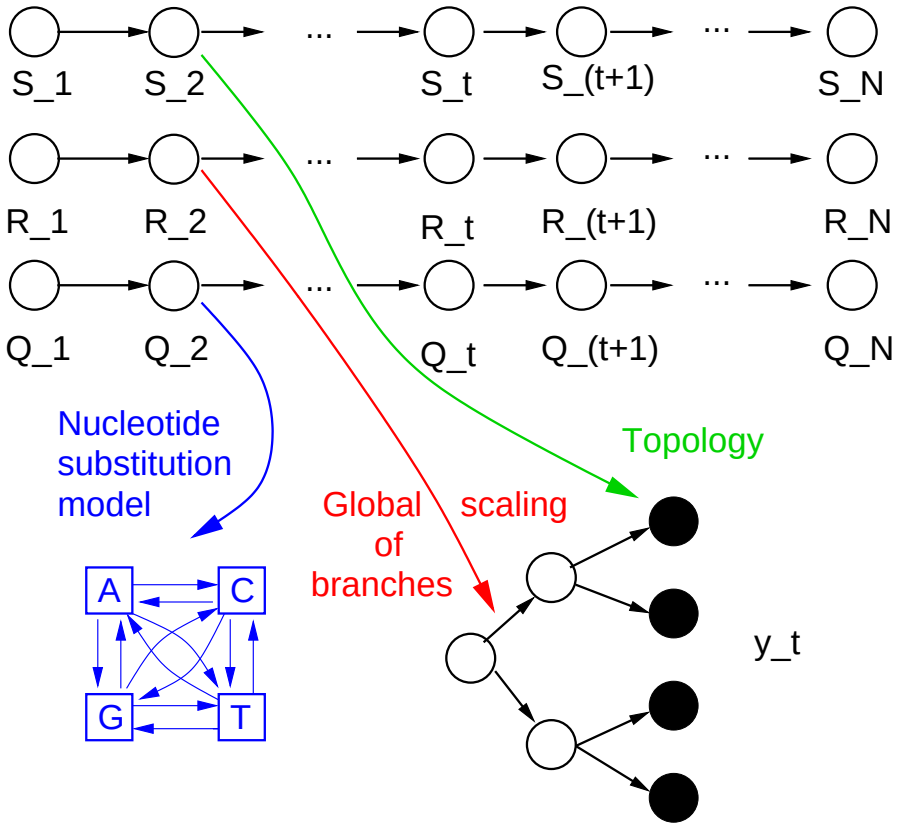


Figure 4.3: **Phylogenetic FHMM.** The figure shows an application of an FHMM, depicted in Figure 4.2, to DNA sequence alignments, with the objective to distinguish between recombination and rate heterogeneity. There are different *a priori* independent hidden Markov chains. Hidden states denoted by S_t represent different phylogenetic tree topologies, and state transitions indicate recombination events. Hidden states denoted by R_t represent different global scaling factors of the branch lengths, and state transitions indicate rate variation. One can optionally include a further chain of hidden states, which represent features of the underlying nucleotide substitution model (here denoted by Q_t).

where the order of these sampling steps is arbitrary. Sampling of the hidden state sequences $\mathbf{S}$ and $\mathbf{R}$ in equations (4.6) and (4.7) is effectively achieved in linear time complexity with the dynamic programming scheme described in Section 3.6 (the stochastic forward-backward algorithm). The distributions for the transition parameters ν_S and ν_R in equations (4.8) and (4.9) are beta distributions, which depend on sufficient statistics computed from the state sequences $\mathbf{R}$ and $\mathbf{S}$, as in equation (3.14).

The distributions of interesting quantities are obtained by marginalization. For instance, the marginal probability of the tree topology at site t, which is used to detect recombination events,

follows from

$$P(S_t|\mathcal{D}) = \sum_{S_1} \dots \sum_{S_{t-1}} \sum_{S_{t+1}} \dots \sum_{S_N} \sum_{\mathbf{R}} \int P(\mathbf{S}, \mathbf{R}, \nu_S, \nu_R|\mathcal{D}) d\nu_S d\nu_R \tag{4.10}$$

$$P(R_t|\mathcal{D}) = \sum_{R_1} \dots \sum_{R_{t-1}} \sum_{R_{t+1}} \dots \sum_{R_N} \sum_{\mathbf{S}} \int P(\mathbf{S}, \mathbf{R}, \nu_S, \nu_R|\mathcal{D}) d\nu_S d\nu_R \tag{4.11}$$

where the marginalization is carried out as described in the text below equation (3.8) and in Section 2.1.

4.2 Findings

Plotting the marginal posterior probabilities for tree topologies $P(S_t|\mathcal{D})$ from (4.10) along the DNA sequence alignment for Neisseria, as in Figure 3.6, eliminates the spurious changepoints associated with rate heterogeneity. These transitions are found when plotting $P(R_t|\mathcal{D})$ from (4.11) along the sequence alignment. For details see [DH20]. In combination with several related studies described in [DH20] these findings demonstrate that the phylogenetic FHMM successfully distinguishes between recombination and rate variation.

4.3 Model Extension

The model proposed in [DH20] associates the rate states R_t with fixed scaling parameters and keeps the number of rate states constant at an *a priori* defined value. I extended this approach in collaboration with Wolfgang Lehrach in [DH9] to infer both the number of rate states and the associated scaling parameters from the posterior distribution. Starting from a truncated Poisson prior on the number of rate states, this trans-dimensional approach is naturally effected with the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm proposed in [54]. The idea is to introduce birth and death moves that change the number of rate states and, hence, the dimension of the associated scaling factor vector $\boldsymbol{\rho}$, while ensuring reversibility of the Markov chain. The increased model flexibility leads to identifiability issues that are common to mixture models, and their solution is discussed in detail in [DH9].

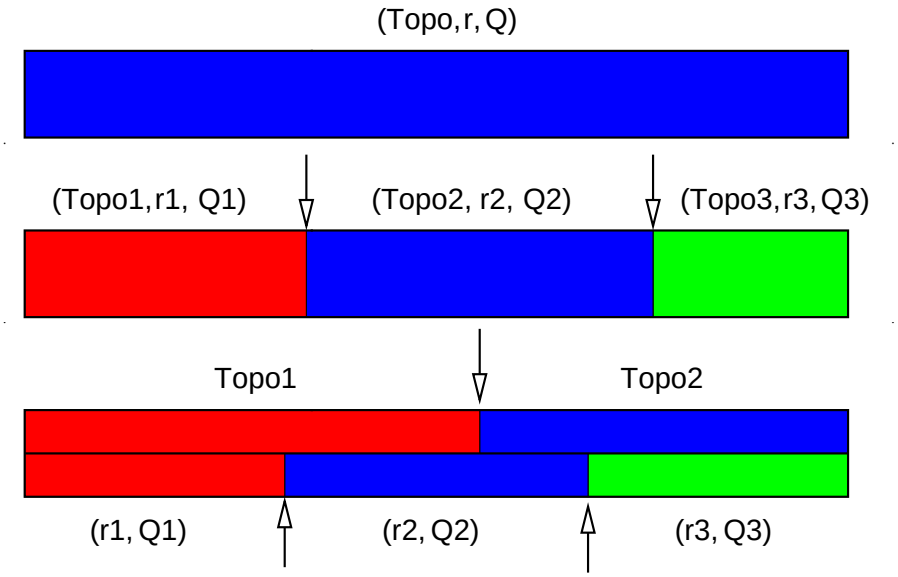


Figure 4.4: **Phylogenetic (dual) multipe changepoint model.** The figure illustrates an alternative to the phylogenetic FHMM, based on (dual) multiple changepoint processes. *Top panel:* A homogeneous DNA sequence alignment. *Centre panel:* The multiple changepoint model, proposed in [139], divides the DNA sequence alignment into separate segments, where each segment is governed by its own phylogenetic tree topology (*Topo*), scaling factor for the branch lengths (r), and nucleotide substitution parameters (Q). *Bottom panel:* The dual multiple changepoint model, proposed in [100], divides the DNA sequence alignment into two different types of segments, where segments of the first type are governed by separate tree topologies (*Topo*), while segments of the second type are associated with different scaling factors for the branch lengths (r) and different nucleotide substitution parameters (Q).

4.4 Comparison with Alternative Approaches from the Literature

The two main competing schemes proposed in the literature are the multiple changepoint model (MCP) of Suchard et al. [139], and the dual multiple changepoint model (DMCP) of Minin et al. [100]. The underlying ideas are illustrated in Figure 4.4. The MCP divides the DNA sequence alignment into segments via a multiple changepoint process, where each segment is governed by its own phylogenetic tree topology and scaling factor for the branch lengths[1] . The number of changepoints is given a truncated Poisson prior, and the changepoints along with all other model parameters are sampled from the posterior distribution with RJMCMC, using birth and death moves for the changepoints. The MCP model is conceptually similar to the phylogenetic HMM, with segments corresponding to states. Like the phylogenetic HMM, it suffers from the inability

[1]Like the phylogenetic FHMM depicted in Figure 4.3, the MCP and DMCP models potentially also allow for changes in the nucleotide substitution parameters. In order to focus this exposition on the essential concepts and avoid extraneous details, I will not discuss this any further here. The interested reader is referred to [100, 139] and [DH9].

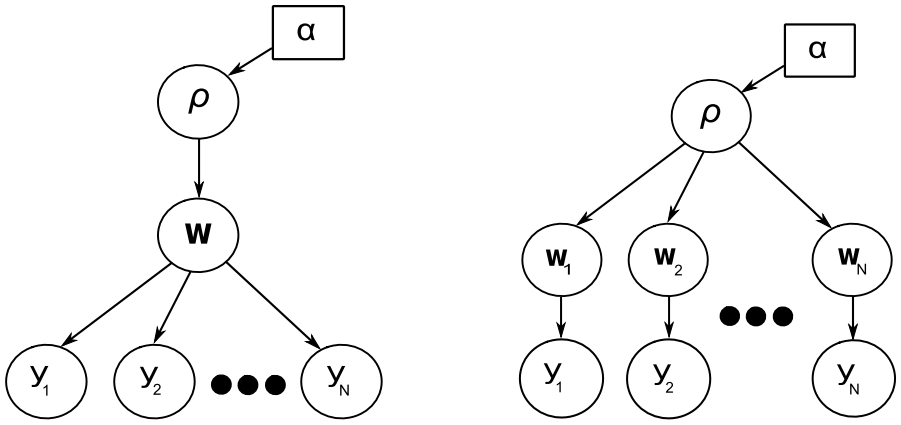
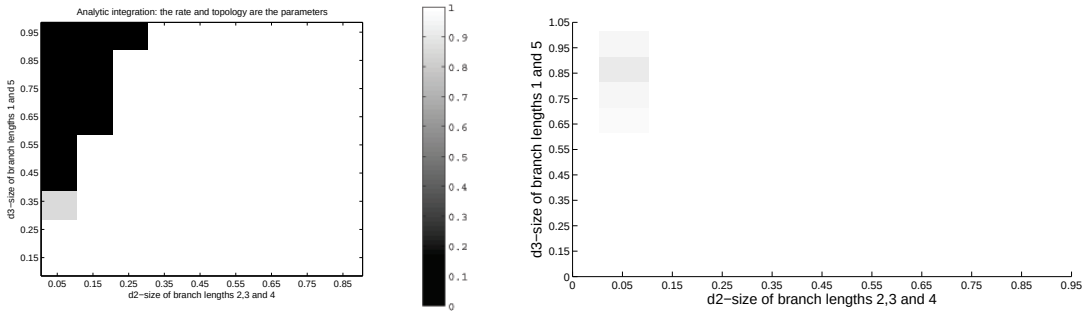


Figure 4.5: **Comparison of the standard phylogenetic model (left) with the no common mechanism model (right).** In both panels, the $\mathbf{y}_t's$ represent the columns in the DNA sequence alignment, $\mathbf{w}$ is a vector of branch lengths, and ρ is a hyperparameter determining the prior distribution over the branch lengths, via equation (4.1), which can itself have a prior distribution depending on some higher-level hyperparameter α. The left panel shows the standard phylogenetic model, with a branch length vector $\mathbf{w}$ common to all sites. The right panel shows the no-common-mechanism model, with separate independent branch length vectors $\mathbf{w}_t$, associated with the sites t in the alignment. As a consequence of this independence assumption, the branch lengths can be integrated out analytically, as described in Section 4.1.

to distinguish between tree topology changes, indicative of recombination, and rate heterogeneity. This shortcoming is addressed with the DMCP, which employs two different types of changepoints: one to mark changes in the phylogenetic tree topology, the other to allow for changes in the global scaling of the branch lengths.

We have compared the DMCP model with our phylogenetic FHMM in [DH9]. The changepoint process can be regarded as a special case of an HMM with unidirectional transition probabilities, meaning that a state can never be visited twice. When a recombinant region is inserted into a DNA sequence alignment, the phylogenetic parameters of the two flanking regions have to be inferred separately, as they constitute different parameter sets. This model is counter-intuitive and potentially leads to an unnecessarily inflated inference uncertainty for short alignments. The problem is avoided with the phylogenetic HMM/FHMM, where the two flanking regions correspond to the same state and therefore share the same parameters.

The second difference is that in the DMCP model the segment lengths and the number of different tree topologies/ differently diverged regions are jointly determined by the distribution over the changepoints. Consequently, these quantities are coupled. In our phylogenetic FHMM, the number of differently diverged regions corresponds to the number of different rate states. This is different from the segment lengths, which are determined by the transition parameter ν_R (the average segment length is $1/(1-\nu_R)$). Consequently, the DMCP is effectively a special case of our phylogenetic FHMM with a fixed value of ν_R and each state occurring only once. In contrast, the



(a) (b) (c)

Figure 4.6: **Demonstration of the failure of the no-common-mechanism model in the Felsenstein zone.** The figures show the posterior probability of the correct tree topology, obtained with Bayesian model selection based on MCMC, as discussed in [DH10]. The tree configurations depend on two branch lengths, $d2$ and $d3$, as defined in Figure 1.8. In each subfigure, the horizontal axis refers to $d2$, and the vertical axis refers to $d3$. The grey shading indicates the value of the inferred posterior probability, as indicated in the legend of Panel (b), ranging from 0 (black) to 1 (white). Panel (a) shows the results obtained for the no-common-mechanism model, represented in the right panel of Figure 4.5. Panel (c) shows the results obtained for the standard model, shown in the left panel of Figure 4.5. The results were obtained from a DNA sequence alignment simulated from the phylogenetic tree of Figure 1.8. The failure of the no-common-mechanism model in the Felsenstein zone becomes evident.

phylogenetic FHMM separates the number of states (or different segment types) from the average segment length, where the average segment length is not set to a fixed value, but is also inferred. For an empirical comparison between the proposed phylogenetic FHMM and the DMCP model, see [DH9].

4.5 Criticism and Relaxation of the No-Common-Mechanism Model

The three models described in the previous sections, the MCP, DMCP and phylogenetic FHMM models, have one feature in common: a completely factorizable prior is placed on the vector of branch lengths, as defined in (4.1) and depicted in the right panel of Figure 4.5. In this way the branch lengths can be integrated out analytically. This is convenient, as the marginal likelihood of the tree topology, the nucleotide substitution rate, and further parameters of the nucleotide substitution model (like the transition-transversion ratio) can be computed in closed from. In this way, the computational complexity of sampling changepoints (MCP,DMCP) or hidden state sequences (PFHMM) from the posterior distribution with MCMC is substantially reduced.

The subject of my joint work with Alexander Mantzaris, reported in [DH10], was to investigate the effect of the approximation on which the analytic integration of the branch lengths is based.

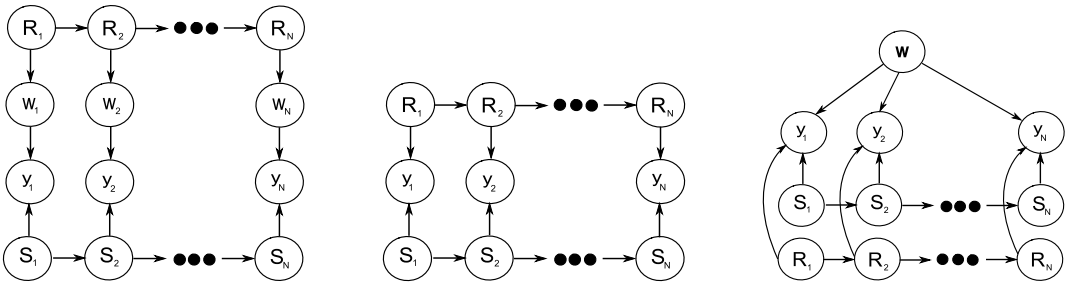


Figure 4.7: **Improvement of the phylogenetic FHMM.** The left panel shows the probabilistic graphical model representation of the phylogenetic FHMM of [DH9,DH20], described in Section 4.1. The $\mathbf{y}_t$'s represent the columns in the DNA sequence alignment, where the subscript $t = 1, \ldots, N$ indicates the site in the alignment. Each site t is associated with a hidden state S_t that defines the tree topology, a vector of branch lengths $\mathbf{w}_t$, and a second hidden state R_t that defines the hyperparameter of the prior distribution on the branch lengths, as defined in equation (4.1). Both hidden states S_t and R_t have a Markovian dependence structure. The chosen form of the model allows the branch lengths to be integrated out analytically, as described in Section 4.1. This results in the simplified model depicted in the centre panel. Note that this model is a phylogenetic factorial HMM, where one type of hidden states $(S_1, \ldots, S_N)$ defines the tree topology, and the other type of hidden states $(R_1, \ldots, R_N)$ defines the average amount of mutational divergence. The right panel shows the probabilistic graphical model representation of the improved phylogenetic factorial HMM proposed in [DH10]. The model is similar to the one presented in the left panels with the essential difference that a common branch length vector $\mathbf{w}$ is shared among all sites. This is a generalization of the standard phylogenetic model, depicted in the left panel of Figure 4.5, which allows for recombination and rate heterogeneity while avoiding the inconsistency of the no-common-mechanism model.

A detailed analysis of the underlying approximation revealed that the resulting model exhibits a behaviour very similar to maximum parsimony, described in Section 1.3. In particular, it suffers from the same inconsistency problem and is intrinsically susceptible to the systematic failure in the Felsenstein zone[2] [40], as we demonstrated with Bayesian model selection. The results are shown in Figure 4.6.

The root of the problem is that for the branch lengths to be integrated out analytically, as in (4.2), the model has to be modified so as to associate a separate branch length vector $\mathbf{w}_t$ with each position t in the DNA sequence alignment. This model is equivalent to the no-common-mechanism model proposed in [145]. It is important to note that it is not the independence assumption of equation (4.1) alone that leads to this simplification, a conclusion one might erroneously draw from [139]. Rather, the more restrictive independence assumption of the no-common-mechanism model, depicted in the right panel of Figure 4.5, is needed. As a consequence of the latter independence assumption the model becomes over-complex, though, with no information sharing between different sites with respect to the branch length estimation. In terms of statistical terminology, the no-common-mechanism model turns the structural parameters $\mathbf{w}$ into a set of incidental parameters[3]. As discussed in [52], this implies that maximum likelihood is no longer guaranteed to provide a

[2]Illustrated in Figure 1.8.

[3]Structural parameters are parameters that appear in the probability distributions of all the observations, whereas an incidental parameter appears in the probability distributions of only a subset of the observations. See [52] for further details.

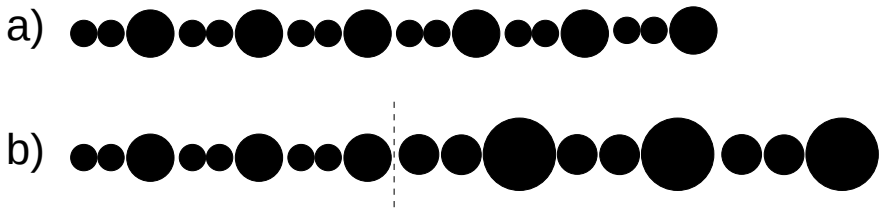


Figure 4.8: Illustration of regional versus within-codon rate heterogeneity. Each circle corresponds to a nucleotide in a DNA sequence, and the circle diameter symbolizes the average nucleotide substitution rate at the respective position. The top panel (a) shows a "homogeneous" DNA sequence composed of six codons, where each third position is more diverged as a consequence of the nature of the genetic code. The bottom panel (b) shows a hypothetical DNA sequence subject to regional rate heterogeneity, where the second half on the right of the dashed vertical line constitutes a region that is more evolved.

consistent estimator. This aspect, which has not been considered for any of the three methods discussed in the previous sections – MCP, DMCP and the factorial FHMM – causes the same inconsistency problems as those found in maximum parsimony.

To address this problem and improve the phylogenetic FHMM of [DH9,DH20], we proposed a modification without the no-common-mechanism assumption for the branch lengths. The difference between the models is depicted in Figure 4.7. This modification increases the computational complexity of the inference scheme, as the branch lengths have now to be numerically sampled from the posterior distribution with MCMC. However, in [DH10] we demonstrate that the resulting model avoids the susceptibility to spurious topology changes in the Felsenstein zone, and thereby improves the accuracy of detecting recombination in DNA sequence alignments.

4.6 Future Work

With four nucleotides occupying three positions in the codon, there are $4^3 = 64$ codons coding for 20 amino acids. This mismatch renders the genetic code intrinsically redundant. The redundancy is such that nucleotide substitutions in the third position occasionally do not change the amino acid, or lead to an amino acid with similar biophysical properties. Consequently, there is less selective pressure on the third codon position, resulting in an effectively increased nucleotide substitution rate. For an illustration, see Figure 4.8.

To prevent any confounding between the rate heterogeneity intrinsic to the genetic code and the regional rate heterogeneity discussed in the previous sections, as illustrated in Figure 4.8, we extended the model so as to disambiguate between these two effects. Define the branch lengths to be of the form $\mathbf{w}_t = r_{R_t}\lambda_{I_t}\tilde{\mathbf{w}}_t$, where the $\tilde{\mathbf{w}}_t$ are normalized such that the L1-norm is equal to 1:

$||\tilde{\mathbf{w}}_t||_1 = 1$. There are two scaling factors, associated with a region (r_{R_t}) and a codon (λ_{I_t}) effect. As before, R_t indicates a rate state of the FHMM. In addition, we have an indicator $I_t \in \{1, 2, 3\}$ for the three codon positions. Note that the normalization constraint $||\tilde{\mathbf{w}}_t||_1 = 1$ can easily be enforced in the MCMC simulations by proposing new vectors $\tilde{\mathbf{w}}_t$ from a Dirichlet distribution. This procedure also ensures that all elements of $\tilde{\mathbf{w}}_t$ are non-negative.

Together with Alexander Mantzaris, I have implemented and tested this model. Results on simulated sequence alignments are available from [DH46]. Applications to real DNA sequence alignments are the subject of his PhD studies.

Part II

Systems Biology

Chapter 5

In Silico Prediction of Protein Interactions

Short well defined domains known as Peptide Recognition Modules (PRMs) regulate many important protein-protein interactions involved in the formation of macromolecular complexes and biochemical pathways. Since high-throughput experiments like yeast two-hybrid and phage display are expensive and intrinsically noisy, it would be desirable to more specifically target or partially bypass them with complementary in silico approaches. In my joint work with Wolfgang Lehrach and Chris Williams [DH19,DH50] we presented a probabilistic discriminative approach to predicting PRM-mediated protein-protein interactions from sequence data. To overcome potential susceptibility to overfitting we adopted a Bayesian *a posteriori* approach based on a Laplacian prior in parameter space. The proposed method was tested on two data sets of protein-protein interactions involving 28 SH3 domain proteins in *Saccharmomyces cerevisiae*, where our discriminative model achieved a better out-of-sample prediction performance than the state-of-the-art probabilistic generative model.

5.1 Overview

Peptide recognition modules (PRMs) are specialized compact protein domains that mediate many important protein-protein interactions. They are responsible for the assembly of critical macromolecular complexes and biochemical pathways [105], and they have been implicated in carcino-

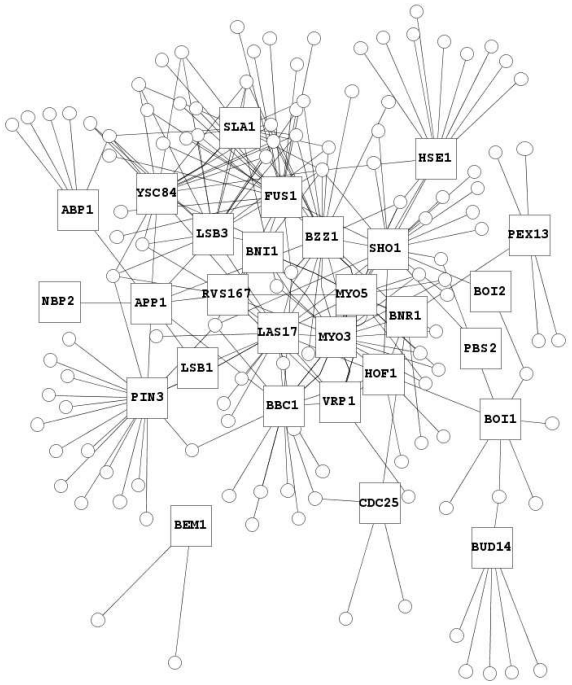


Figure 5.1: **The yeast two hybrid interaction network of SH3.** The labelled squares represent the central SH3 domains, while the circles represent the peripheral proteins that were found to bind to the SH3 domains.

genesis and various other human diseases [140]. PRMs recognize and bind to peptide ligands that contain a specific structural motif. One of the most actively studied PRMs is the SH3 domain, which binds to peptide ligands that contain a particular proline-rich core. Tong et al. [144] carried out two extensive experimental studies to infer the network of SH3-mediated protein-protein interactions in *Saccharmomyces cerevisiae*. They identified 28 SH3 domain proteins in the *S. cerevisiae* proteome, which were used as baits and screened against conventional and proline-rich libraries in a yeast two-hybrid experiment [146]. In a second independent study, they screened random peptide libraries by phage display [146] to identify the consensus sequence for preferred ligands that bind to each PRM. Based on these consensus sequences, they inferred a protein-protein interaction network that links each PRM to proteins containing the preferred ligand. Since both experimental procedures are intrinsically noisy, the two independently inferred interaction networks were found to show only a modest degree of overlap, as shown in Figure 5.2. Reiss and Schwikowski [111] addressed the question of whether computational *in silico* approaches would allow some of the difficult and expensive experimental procedures to be more specifically targeted, or even bypassed altogether. To this end, they developed a probabilistic generative model of the SH3 ligand peptides, based on the widely used Gibbs sampling motif finding algorithm [81, 88]. Directly applying the standard Gibbs motif sampler to the *S. cerevisiae* SH3 interaction data faces the difficulty that each SH3 domain is only involved in a small number of interactions (between 1 and 20), which leads to a poor motif conservation and a high susceptibility to random artifacts due to the small

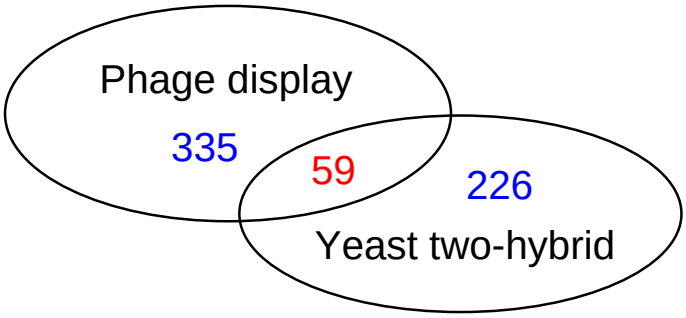


Figure 5.2: **Protein interactions obtained with different experimental procedures.** Tong et al. [144] compared two different assays for measuring protein interactions related to 28 SH3 domain proteins in yeast. With yeast two-hybrid, they found 285 interactions. With phage display, they found 394 interactions, involving the same SH3 domain proteins. Only 59 interactions were consistently detected with both assays.

sample size. Conversely, searching for a single motif in all identified SH3 domains lacks the specificity to identify anything but a broad consensus pattern. Reiss and Schwikowski [111] therefore devised a compromise strategy, where the network information was used as a prior on the structure of individual motifs, which were searched for with a modified version of the Gibbs motif sampler. The prior was adjusted to become discriminative, giving higher probability to those motifs that are distinct from non-binding motifs.

Reiss and Schwikowski [111] encouragingly demonstrate that a probabilistic model trained on protein sequences and observed physical interactions can succeed in independently predicting new protein-protein interactions mediated by SH3 domains. However, a shortcoming of their model is a dependence on tuning parameters that have to be chosen in advance by the user and that are not inferred from the data. Inappropriate values reduce the performance of their algorithm to using standard motif searching algorithms, and it is unlikely that universal values applicable to different protein (super-) families exist. Also, the proposed model borrows substantial strength from its heuristic discriminative modification of the prior, which again depends on various tuning parameters.

In my joint work with Wolfgang Lehrach and Chris Williams [DH19,DH50] we proposed an alternative *in silico* method for the prediction of SH3-mediated protein-protein interactions, which addresses some of the shortcomings of the model introduced in [111]. A key feature of our model is that it is discriminative: given a set of protein sequences, the model only attempts to find domains that distinguish between different SH3 binding domains. This is in contrast to the approach in [111], which is based on a generative model of the whole sequence. As discussed in [130], a generative approach can be confounded by repetitive or over-represented motifs that are unrelated to PRM-peptide interactions, which our discriminative model avoids by formulating the learning problem in terms of a supervised classification problem.

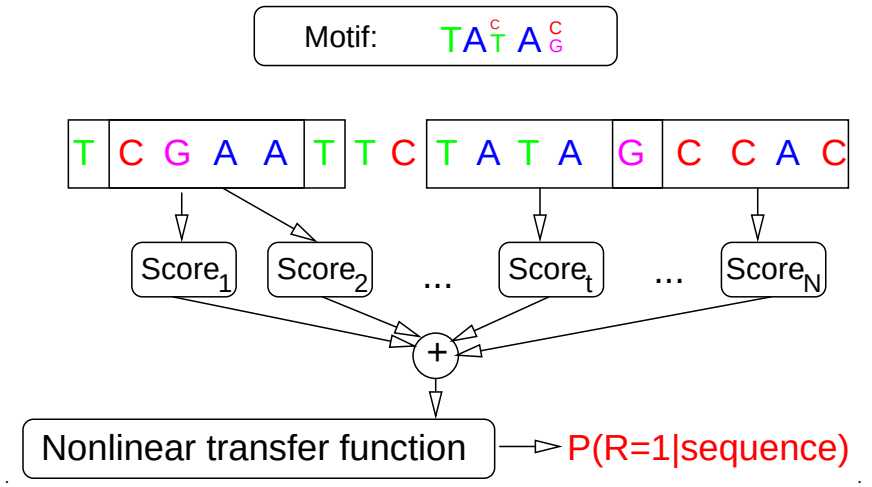


Figure 5.3: **Illustration of the discriminative model for predicting protein-peptide interactions, as defined by equation (5.1).** The inner sum in (5.1) is a motif detector, the outer sum extends over all possible motif positions, and the result of the summation is put through a nonlinear transfer function of the form logit(log(...)). A derivation can be found in [DH19].

The model we propose is based on a DNA-sequence model applied in [129, 130]. However, due to the larger size of the alphabet (20 amino acids instead of 4 nucleotides) and the small number of interactions per SH3 domain, their maximum likelihood approach to parameter estimation is bound to lead to serious overfitting. An essential component of our approach, therefore, is the inclusion of a regularization scheme, resulting in a maximum a posteriori (MAP) or penalized maximum likelihood estimate of the parameters based on a Laplacian (L1 norm) prior. The methodological details can be found in [DH19].

5.2 Method

This section provides a brief sketch of the method; for a proper exposition, see [DH19]. Denote by $R \in \{0, 1\}$ a binary variable indicating the presence ($R = 1$) or absence ($R = 0$) of a protein-peptide interaction, and by $\mathbf{x} = [S_1, S_2, \ldots, S_N]$ a peptide sequence of N amino acids, with S_t denoting the amino acids at the tth position in the sequence. Given this sequence, the probability of an interaction with a specified peptide recognition module is given by

$$P(R = 1|S_1, S_2, \ldots, S_N) \quad = \quad \text{logit}\left(\log \left[\frac{w_0}{N - W + 1} \sum_{t=0}^{N-W} \exp \left(\sum_{k=1}^{W} w_k(S_{t+k}) \right) \right] \right) \qquad (5.1)$$

The model has $20W + 1$ free parameters: $\mathbf{w} = \{w(0), w_k(1), \ldots, w_k(20)\}; k \in \{1, 2, \ldots, W\}$, where W is an *a priori* chosen constant that indicates the expected length of the binding motifs, and $w_k(a)$ denotes a weight associated with intra-motif position k for amino acid $a \in \{1, \ldots, 20\}$ (recall that there are 20 amino acids). An illustration is given in Figure 5.3, and a proper derivation of (5.1) can be found in [DH19]. A brief sketch of the underlying idea is as follows. Assume that a sequence without any binding motif, $R = 0$, is distributed according to the background distribution

$$P(S_1, S_2, \ldots, S_N | R = 0) \; = \; \prod_{t=1}^{N} \theta_0(S_t) \tag{5.2}$$

with $\theta_0(i) \in [0, 1]$ for all amino acids $i \in \{1, \ldots, 20\}$, and $\sum_{i=1}^{20} \theta_0(i) = 1$. Given a binding sequence ($R = 1$), with a binding motif of length W starting at position $m + 1$, the probability of the respective sequence is

$$\begin{aligned} & P(S_1, S_2, \ldots, S_N | R = 1, start = m + 1) \\ & = \; \prod_{t=1}^{m} \theta_0(S_t) \prod_{k=1}^{W} \psi_k(S_{m+k}) \prod_{t=m+W+1}^{N} \theta_0(S_t) \; = \; \prod_{t=1}^{N} \theta_0(S_t) \prod_{k=1}^{W} \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})} \end{aligned} \tag{5.3}$$

The parameters $\psi_k(i)$, $1 \leq k \leq W$ and $1 \leq i \leq 20$, define the so-called position sensitive scoring matrix of the binding motif, and we have for all $1 \leq k \leq W$: $\psi_k(i) \in [0, 1]$, $\sum_{i=1}^{20} \psi_k(i) = 1$. The probability of a binding sequence ($R = 1$) with a motif starting anywhere is given by marginalization:

$$\begin{aligned} & P(S_1, S_2, \ldots, S_N | R = 1) \\ & = \; \sum_{m=0}^{N-W} P(start = m + 1) P(S_1, S_2, \ldots, S_N | R = 1, start = m + 1) \\ & = \; \left(\prod_{t=1}^{N} \theta_0(S_t) \right) \frac{1}{N - W + 1} \sum_{m=0}^{N-W} \prod_{k=1}^{W} \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})} \end{aligned} \tag{5.4}$$

where a uniform prior on the binding locations, $P(start = m + 1)$, has been assumed. We now apply Bayes rule to obtain:

$$\begin{aligned} P(R = 1 | S_1, S_2, \ldots, S_N) \quad & = \quad \frac{P(S_1, S_2, \ldots, S_N | R = 1) P(R = 1)}{P(S_1, S_2, \ldots, S_N)} \\ & = \; \frac{P(S_1, S_2, \ldots, S_N | R = 1) P(R = 1)}{P(S_1, S_2, \ldots, S_N | R = 0) P(R = 0) + P(S_1, S_2, \ldots, S_N | R = 1) P(R = 1)} \\ & = \; \left(1 + \frac{P(R = 0) P(S_1, S_2, \ldots, S_N | R = 0)}{P(R = 1) P(S_1, S_2, \ldots, S_N | R = 1)} \right)^{-1} \\ & = \; \left(1 + \left[\frac{P(R = 1)}{P(R = 0)} \frac{1}{(N - W + 1)} \sum_{m=0}^{N-W} \prod_{k=1}^{W} \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})} \right]^{-1} \right)^{-1} \end{aligned}$$

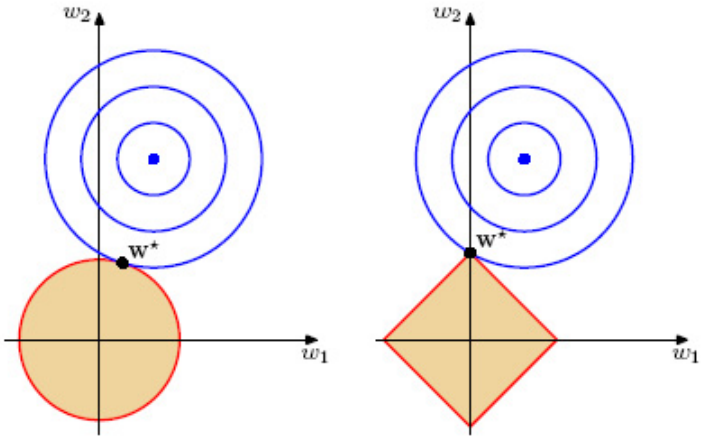


Figure 5.4: **Motivation for the Laplace prior.** The figure illustrates the contours of the log likelihood (blue) along with the log prior (shaded area) for the Gaussian prior on the left and the Laplacian prior (Lasso) on the right. The maximum *a posteriori* (MAP) parameter vector $\mathbf{w}$ is denoted by $\mathbf{w}^*$. It is seen that the Lasso gives a sparser solution in which one of the parameters (w_2) is exactly zero. Illustration taken from [14].

With the definitions

$$w_k(l) = \log \frac{\psi_k(l)}{\theta_0(l)}, \quad w_0 = \frac{P(R=1)}{P(R=0)}, \quad \text{logit}(z) = \frac{1}{1+\exp(-z)} \tag{5.5}$$

we get equation (5.1).

Given a training set $\mathcal{D} = \{\mathbf{x}_i, R_i\}; 1 \leq i \leq T$ of T peptide sequences $\mathbf{x}_i$ and their associated binding indicators R_i, the (log) likelihood is given by

$$\begin{aligned} P(\mathcal{D}|\mathbf{w}) &= \prod_{i=1}^{T} y(\mathbf{x}_i, \mathbf{w})^{R_i} [1 - y(\mathbf{x}_i, \mathbf{w})]^{(1-R_i)} \\ \log P(\mathcal{D}|\mathbf{w}) &= \sum_{i=1}^{T} R_i \log y(\mathbf{x}_i, \mathbf{w}) + (1 - R_i) \log y(\mathbf{x}_i, \mathbf{w}) \end{aligned} \tag{5.6}$$

where $y(\mathbf{x}_i, \mathbf{w})$ is given by (5.1). As the maximum likelihood configuration is susceptible to overfitting, we include a regularization term based on the L1-norm (LASSO), proposed in [142, 152]. This corresponds to a Laplacian prior, $P(\mathbf{w}) \propto \exp(-\alpha|\mathbf{w}|)$, and we aim to estimate the parameter vector $\mathbf{w}$ in a MAP (maximum *a posteriori*) sense:

$$\mathbf{w}^* = \text{argmax}_{\mathbf{w}} P(\mathbf{w}|\mathcal{D}) = \text{argmax}_{\mathbf{w}} \left[\log P(\mathcal{D}|\mathbf{w}) + \log P(\mathbf{w})\right] \tag{5.7}$$

The motivation for this prior is as follows. First, it shrinks the parameters to zero. From the definitions (5.5) it is seen that zero parameters correspond to the absence of an interaction; hence

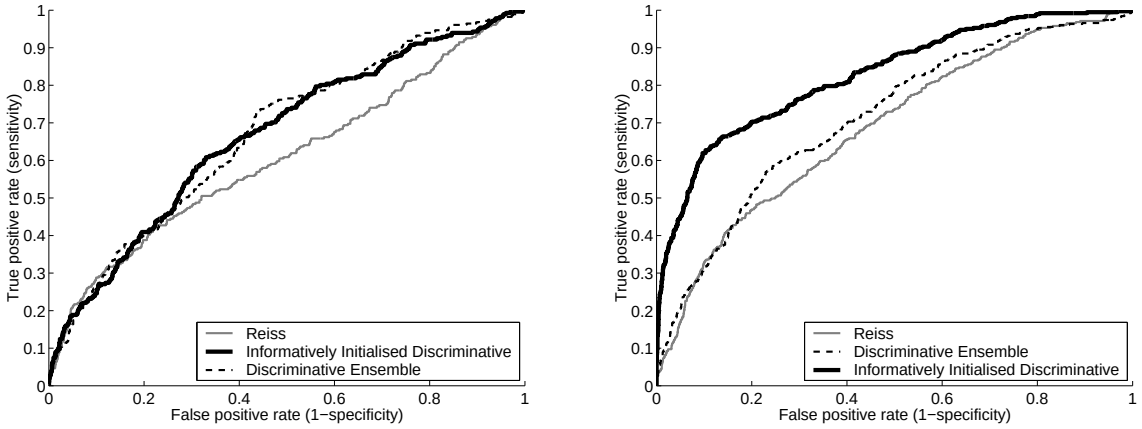


Figure 5.5: **Comparative evaluation of *in silico* methods for predicting protein-peptide interactions.** The graphs show ROC curves obtained for the three methods compared in our study: GEN (narrow solid lines): generative model of Reiss and Schwikowski [111]; DIS-I (thick solid line): discriminative model, informative initialization; and DIS-E (dashed line): ensemble of discriminative models, random initializations. A ten-fold cross-validation scheme was applied, as described in the main text. *Left panel:* yeast two-hybrid; *right panel:* phage display. The areas under the curves are, for yeast two-hybrid: GEN: 0.61, DIS-I: 0.67, DIS-E:0.67, and for phage display: GEN: 0.69, DIS-I: 0.83, DIS-E:0.71.

the prior has a regularization effect that discourages spurious interactions. This regularization effect is stronger for the Laplacian prior (Lasso) than the Gaussian prior (ridge regression) in that more parameters tend to be driven down to zero, leading to sparser network structures. An illustration is given in Figure 5.4. Since there is no closed-form solution to this optimization problem (5.7), we follow a gradient ascent scheme. The details, including the setting of the hyperparameter α, are discussed in [DH19].

5.3 Findings

We applied the proposed discriminative model with different regularization schemes to the phage display and yeast two-hybrid protein interaction data of Tong et al. [144].[1] We evaluated the generalization performance with a 10-fold cross-validation scheme where the data were randomly partitioned into 10 folds. The generalization performance was then evaluated on the current fold, and the other 9 folds were used for training. We obtained an average out-of-sample performance

[1]We removed SH3 domain proteins that only bind to a single peptide, as there would be no way to validate these interactions on an independent test set. With this modification, the phage display data set contains 17 SH3 domains, 207 binding partners, and 381 interactions, while the yeast two-hybrid data set (displayed in Figure 5.1) has 28 SH3 domains, 143 binding partners, and 285 interactions. Further details can be found in the supplementary material of [144].

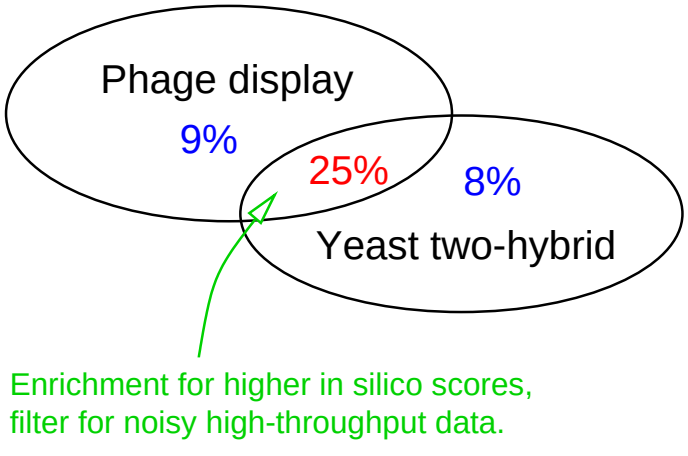


Figure 5.6: **Biological validation of the predicted protein-peptide interaction.** When training our model on the yeast two-hybrid data, we recovered 25 percent of the interactions in the intersection set, but only 8 percent of the interactions in the complementary non-intersection set. When training our model on the phage display data, we again recovered 25 percent of the interactions in the intersection set, but only 9 percent of the interactions in the non-intersection set. This indicates that the subset of more reliable interactions (i.e. interactions found with both assays) is noticeably enriched for high-scoring *in silico* predictions.

by repeating this for all 10 folds.

For comparison with [111], we measured the performance in terms of receiver operating characteristics (ROC) curves, which are explained in Section 6.2. In brief, they are obtained by subjecting the predicted posterior probabilities of protein-peptide interactions to a threshold parameter $\theta \in [0, 1]$, keeping only those interactions whose posterior probability exceeds θ, and plotting the proportion of predicted true interactions against the proportion of incurred false interactions for all values $\theta \in [0, 1]$. By numerically integrating over the whole range of $\theta \in [0, 1]$ we obtain the area under the curve (AUC). This AUC score ranges from 0.5 for a random predictor to 1.0 for a perfect predictor, with larger values generally indicating a better performance.

In our simulations we found that the Laplacian (L1-norm) regularization scheme achieved a significant improvement (in terms of out-of-sample prediction accuracy) on both the unregularized as well as the Gaussian (L2-norm) regularized models. We further compared three methods: (1) the generative model of Reiss and Schwikowski [111]; (2) the proposed discriminative model, where the weights were initialized from the position weight matrices of the sequence motifs learned with the generative model; and (3) an ensemble of discriminative models; this ensemble was created by training ten models from different initializations, and keeping the five models with the highest training set scores. In what follows, I will refer to these methods as (1) GEN, (2) DIS-I, and (3) DIS-E, respectively. For training the discriminative models in the comparative evaluation, the Laplacian regularization scheme was applied throughout.

The left panel of Figure 5.5 shows the ROC curves obtained for the yeast two-hybrid network. Both discriminative methods, DIS-I and DIS-E, clearly outperform GEN in the right part of the graph, for false positive rates (FPR) greater than 0.3. This is reflected in higher overall AUC scores, as shown in the caption of Figure 5.5. The right panel of Figure 5.5 shows the ROC curves obtained for the phage display network. The discriminative methods outperform the generative model, and this improvement is considerably improved when starting the training simulations from an informative initialization (DIS-I). We also found that the performance of all methods is consistently better for the phage display network than for the yeast two-hybrid network, as the former assay is more geared towards interactions that are mediated by short sequence binding motifs, i.e. more in line with our modelling approach. We have discussed the detection and localization of these binding motifs in [DH19].

An important practical application of the proposed method is the cleaning and filtering of high-throughput interaction data. Our conjecture was that protein interactions that are assigned a higher posterior probability score *in silico* are more reliable than those with a lower score. We therefore assumed that interactions found with both the yeast two-hybrid as well as the phage display experiments have, on average, higher posterior probability scores than those found with only one experiment. Phrased differently, we assumed that the intersection of the sets of interactions found with yeast two-hybrid and phage display would show an enrichment for higher-scoring *in silico* interactions. To test this conjecture, we extracted for both experiments the 400 highest-scoring interactions; this is the number of interactions detected experimentally with phage display. When training our model on the yeast two-hybrid data, we recovered 25 percent of the interactions in the intersection set, but only 8 percent of the interactions in the complementary non-intersection set. When training our model on the phage display data, we again recovered 25 percent of the interactions in the intersection set, but only 9 percent of the interactions in the non-intersection set. An illustration is given in Figure 5.6. Hence, in both training simulations, we found that the subset of more reliable interactions (that is, those interactions found with both experimental methods) was noticeably enriched for high-scoring *in silico* predictions. This finding corroborated our hypothesis that the predicted interaction scores are biologically consistent and that our method thus offers a useful tool for filtering noisy high-throughput protein interaction data.

Chapter 6

Reverse Engineering Gene Regulatory Networks

Molecular pathways consisting of interacting proteins underlie the major functions of living cells, and a central goal of systems biology is the elucidation of their structure and regulatory mechanisms. Summarizing my work in [DH3,DH4,DH5,DH12,DH13,DH14,DH16,DH18,DH23,DH41,DH42,DH44,DH45], this chapter covers the reconstruction of gene regulatory networks from transcriptomic profiles with probabilistic graphical models. It discusses the evaluation of the network reconstruction accuracy from a realistic simulation study [DH18,DH23], the systematic integration of prior knowledge in a Bayesian framework [DH13,DH16], the inference of evolving network structures during an organism's life cycle [DH41,DH42], the approximate modelling of nonlinear regulation [DH12,DH44,DH45], and improved MCMC schemes [DH3,DH4,DH14] for Bayesian learning of Bayesian networks.

6.1 Introduction

Molecular pathways consisting of interacting proteins underlie the major functions of living cells. A central goal of molecular biology is therefore to understand the regulatory mechanisms of gene transcription and protein synthesis, and the invention of DNA microarrays and deep sequencing technologies, by which the transcription levels of thousands of genes can be measured simultaneously, mark an important breakthrough in this endeavour. Several approaches to the reverse engineering of genetic regulatory networks from gene expression data have been explored, reviewed,

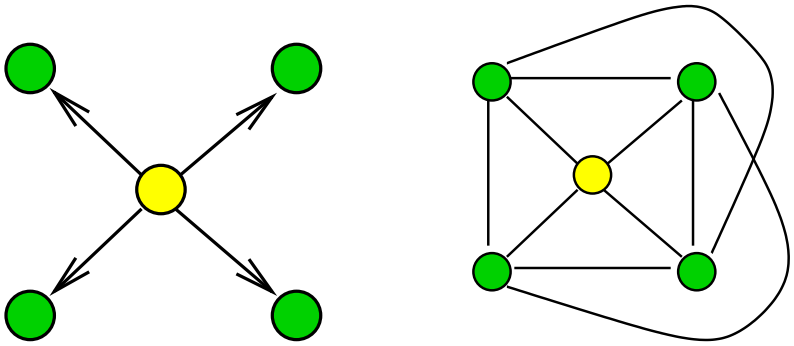


Figure 6.1: **Shortcomings of relevance networks.** The left figure shows a hypothetical gene regulatory network, where the transcription of the centre gene (the 'regulator') initiates the transcription of four other genes (the 'regulatees'). The right figure shows the outcome from the application of relevance networks. Since all regulated genes are under the influence of the same regulator, their expression profiles tend to be correlated and to have high mutual information scores. All five genes are consequently grouped together, and the actual nature of the regulation scheme remains unknown. Note that any approach based on clustering effectively leads to the same result, and hence suffers from the same shortcoming.

for example, in [32, 34, 151].

At the most refined level of detail is a mathematical description of the biophysical processes in terms of a system of coupled differential equations that describe, for example, the processes of transcription factor binding, diffusion, and RNA degradation; see, for instance, [26, 151]. While such low-level dynamics are critical to a complete understanding of regulatory networks, they require detailed specifications of both the relationship between the interacting entities as well as the parameters of the biochemical reaction, such as reaction rates and diffusion constants. Zak et al. [156] found that a system of ordinary differential equations describing a regulatory network of three genes with their respective mRNA and protein products is not identifiable when only gene expression data are observed, and that rich data, including detailed information on protein–DNA interactions, are needed to ensure identifiability of the parameters that determine the interaction structure. In a more recent study, Vyshemirsky and Girolami [149] successfully demonstrated the computation of the marginal likelihood for Bayesian ranking of biochemical system models described by systems of coupled differential equations. Calderhead et al. [23] found a way to reduce the computational costs with a nonlinear regression approach based on Gaussian processes to avoid the need for an explicit solution of the differential equation. However, even with this trick, the computational costs are substantial. Hence while the ideas presented in [23, 149] provide a powerful approach to model selection and hypothesis testing, they are not viable for the *ab initio* reconstruction of regulatory networks.

At the other extreme of the spectrum is the coarse-scale approach of clustering, illustrated in Figure 6.1. Following up on the seminal paper by Eisen et al. [38], several clustering methods have been applied to gene expression data, reviewed, for instance, in [34]. Clustering provides a

computationally cheap way to extract useful information out of large-scale expression data sets. The underlying conjecture is that co-expression is indicative of co-regulation; thus clustering may identify genes that have similar functions or are involved in related biological processes. The disadvantage, however, is that clustering only indicates which genes are co-regulated; it does *not* lead to a fine resolution of the interaction processes that indicates, for example, whether an interaction between two genes is direct or mediated by other genes, whether a gene is a regulator or regulatee, and so on. Clustering, in effect, only groups interacting genes together in a monolithic block, where the detailed form of the regulatory interaction patterns is lost. In the same vein, the popular and widely applied approach of relevance networks, which produces 'networks' by connecting genes with significant correlation or mutual information scores [21, 22], does not distinguish between direct and indirect interactions either and therefore intrinsically fails to produce any *meaningful* regulatory network structure; see Figure 6.1 for an illustration.

A promising compromise between these two extremes is the approach of Bayesian networks. Bayesian networks have received increasing attention from the computational biology community as models of gene regulatory networks, following up on pioneering work by Friedman et al.[46] and Hartemink et al.[57]. Several tutorials on Bayesian networks have been published [60, 77], including a comprehensive tutorial by myself; see Chapter 2 in [DH1]. I will therefore only qualitatively recapitulate some aspects that are of relevance for the present chapter, and refer the reader to the above tutorials for a thorough and more rigorous introduction.

The structure of a Bayesian network is defined by a directed acyclic graph (DAG) indicating how different variables of interest, represented by nodes,"interact". The word "interact" has a causal connotation, which is ultimately of interest to the biologist, but has to be taken with caution in this context, as explained shortly. The edges of a Bayesian network are associated with conditional probabilities, defined by a functional family and their parameters. The interacting entities are associated with random variables, which represent some measured quantities of interest, like relative gene expression levels or protein concentrations. We denote the set of all the measurements of all the random variables as the data, represented by the letter $\mathcal{D}$. As a consequence of the acyclicity of the network structure, the joint probability of all the random variables can be factorized into a product of lower-complexity conditional probabilities according to conditional independence relations defined by the graph structure $\mathcal{M}$. The conditional probabilities depend on some parameters $\mathbf{q}$, which characterize the interactions between the nodes. Under certain regularity conditions, the parameters associated with these conditional probabilities can be integrated out analytically:

$$P(\mathcal{D}|\mathcal{M}) \;=\; \int P(\mathcal{D}|\mathcal{M},\mathbf{q})P(\mathbf{q}|\mathcal{M})d\mathbf{q} \tag{6.1}$$

This allows us to compute the marginal likelihood or evidence $P(\mathcal{D}|\mathcal{M})$, which captures how well the network structure $\mathcal{M}$ explains the data $\mathcal{D}$. Two widely applied models for the likelihood $P(\mathcal{D}|\mathcal{M},\mathbf{q})$ and the prior $P(\mathbf{q}|\mathcal{M})$ are the multinomial distribution (for discretized data) with a conjugate Dirichlet prior, and a linear Gaussian distribution (for continuous data) with a conjugate normal-Wishart prior. The resulting score $P(\mathcal{D}|\mathcal{M})$ was derived in [48, 61, 62] and is referred to

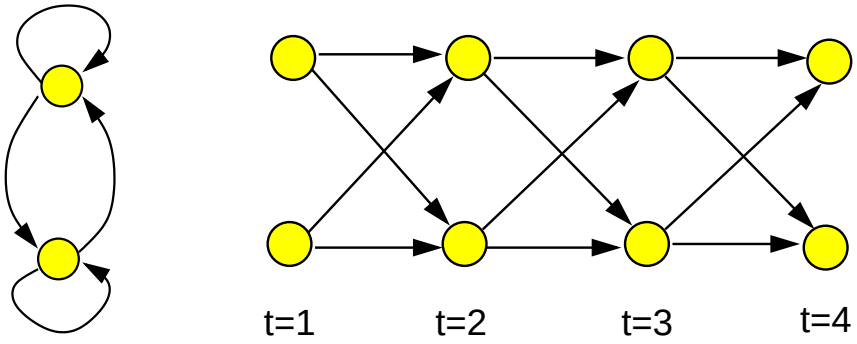


Figure 6.2: **Modelling feedback with dynamic Bayesian networks.** *Left:* Recurrent network comprising two genes with feedback that interact with each other. This structure is *not* a Bayesian network. *Right:* Equivalent dynamic Bayesian network obtained by unfolding the recurrent network in time. Note that similar unfolding methods have been applied in the study of recurrent neural networks ([65], page 183).

as the BDe score (discrete data) or BGe score (continuous data), respectively.

We are interested in learning a network of causal relations between interacting nodes. While such a causal network forms a valid Bayesian network, the inverse relation does not hold: when we have learned a Bayesian network from the data, the resulting graph does not necessarily represent the correct causal graph. One reason for this discrepancy is the existence of unobserved nodes. When we find a probabilistic dependence between two nodes, we cannot necessarily conclude that there exists a causal interaction between them, as this dependence could have been brought about by a common yet unobserved regulator. However, even under the assumption of complete observation the inference of causal interaction networks is impeded by symmetries within so-called equivalence classes, which consist of networks that yield the same evidence scores $P(\mathcal{D}|\mathcal{M})$. A simple example are two conditionally dependent nodes, say A and B, where the two networks related to the two possible directions of the edge, $A \rightarrow B$ and $A \leftarrow B$, are equivalent.

There are three ways to break the symmetries of the equivalence classes. One approach is to use active interventions, like gene knockouts and over-expressions. When knocking out gene A affects gene B, while knocking out gene B does not affect gene A, then $A \rightarrow B$ will tend to have a higher evidence than $A \leftarrow B$. For more details, see [107] and [DH18]. An alternative way to break the symmetries, investigated in more detail in Section 6.3, is to use prior information. When genes A and B are conditionally dependent, and we have prior knowledge that A is a transcription factor that regulates genes in the functional category that B belongs to, then we will presumably favour $A \rightarrow B$ over $A \leftarrow B$. To formalize this notion, we score networks by the posterior probability

$$P(\mathcal{M}|\mathcal{D}) \;\propto\; P(\mathcal{D}|\mathcal{M})P(\mathcal{M}) \tag{6.2}$$

where $P(\mathcal{D}|\mathcal{M})$ is the evidence, and $P(\mathcal{M})$ is the prior distribution over network structures; the latter distribution captures the biological knowledge that we have prior to measuring the data $\mathcal{D}$. While different graphs might have identical scores in light of the data, $P(\mathcal{D}|\mathcal{M})$, symmetries

can be broken by the inclusion of prior knowledge, $P(\mathcal{M})$, and these two sources of information are systematically integrated into the posterior distribution $P(\mathcal{M}|\mathcal{D})$. A third approach is to use temporal information, which intrinsically breaks the symmetry between cause and effect (as the former has to precede the latter). Under the Markov assumption the process of unfolding the network in time leads to a so-called dynamic Bayesian network (DBN), as illustrated in Figure 6.2. This unfolding process also overcomes the acyclicity restriction of static Bayesian networks and allows the modelling of feedback loops and recurrent structures.

The objective of inference is to find the network structure $\mathcal{M}$ that maximizes $P(\mathcal{M}|\mathcal{D})$. Unfortunately, the number of structures increases super-exponentially with the number of nodes. Also, in systems biology, where we aim to learn complex interaction patterns involving many components, the amount of information from the data and the prior is usually not sufficient to render the distribution $P(\mathcal{M}|\mathcal{D})$ sharply peaked at a single graph. Instead, the distribution is usually diffusely spread over a large set of networks. Summarizing this distribution by a single network would not be appropriate. Instead, we aim to sample network structures from the posterior distribution $P(\mathcal{M}|\mathcal{D})$ so as to obtain a typical collection of high-scoring networks and, thereby, capture intrinsic inference uncertainty. Direct sampling from this distribution is usually intractable, though. Hence, we resort to a Markov chain Monte Carlo (MCMC) scheme [90], which under fairly general regularity conditions is theoretically guaranteed to converge to the posterior distribution of equation (6.2). Given a network structure $\mathcal{M}_{\text{old}}$, a new network structure $\mathcal{M}_{\text{new}}$ is proposed from a proposal distribution $Q(\mathcal{M}_{\text{new}}|\mathcal{M}_{\text{old}})$, which is then rejected or accepted according to the standard Metropolis-Hastings scheme [59] with the following acceptance probability:

$$A = \min\left\{\frac{P(\mathcal{D}|\mathcal{M}_{\text{new}})P(\mathcal{M}_{\text{new}})}{P(\mathcal{D}|\mathcal{M}_{\text{old}})P(\mathcal{M}_{\text{old}})}\frac{Q(\mathcal{M}_{\text{old}}|\mathcal{M}_{\text{new}})}{Q(\mathcal{M}_{\text{new}}|\mathcal{M}_{\text{old}})}, 1\right\} \tag{6.3}$$

The functional form of the proposal distribution $Q(\mathcal{M}_{\text{new}}|\mathcal{M}_{\text{old}})$ depends on the chosen type of proposal moves. For a derivation, see [DH1], Chapter 2. While edge-based operations are conceptually easy, the resulting changes in structures space are small and, consequently, convergence and mixing of the Markov chains is usually poor. Together with Marco Grzegorczyk I have developed a new proposal scheme, which enables larger changes in the network structure at high acceptance probability. We have demonstrated that our modification leads to a considerable improvement in the mixing and convergence of the Markov chains. The details can be found in [DH14].

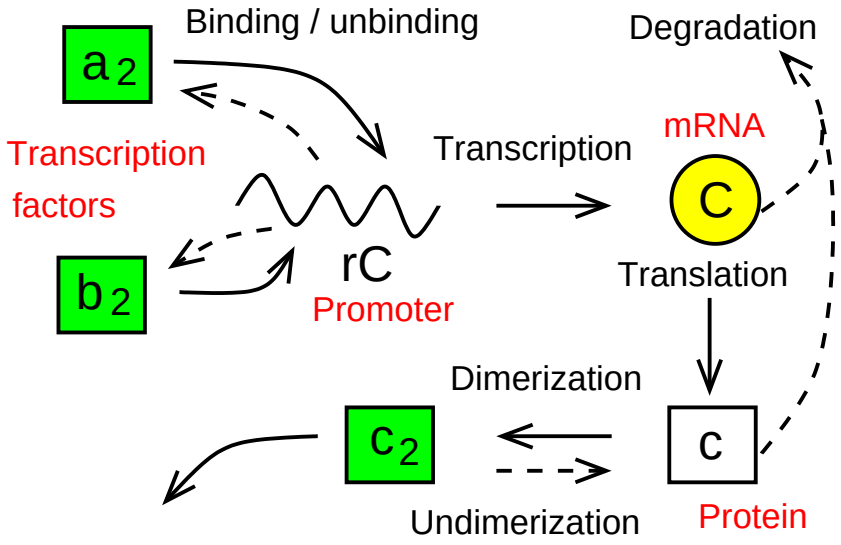


Figure 6.3: **Elementary molecular biological processes.** Two transcription factor dimers a_2 and b_2 bind to the *cis* regulatory site rC in the promoter region upstream of a gene, influencing its rate of transcription. The transcribed mRNA C is translated into protein c, which dimerizes into c_2 to form a new active transcription factor that can bind to other *cis* regulatory sites.

6.2 Assessing the Accuracy of Network Reconstruction

I have assessed the accuracy of gene regulatory network inference based on a simulation study with a known true network and data-generating processes similar to those found in real biological systems. My study, which is described in [DH23], was based on the regulatory network proposed by Zak et al. [155], which is shown in Figure 6.4 and which contains several structures similar to those reported in the literature, like a hysteretic oscillator [8], a genetic switch [47], as well as a ligand binding mechanism that influences transcription. The elementary processes are shown in Figure 6.3 and are described by the following system of differential equations, which describe the processes of transcription factor binding, transcription, translation, dimerization, mRNA degradation, and protein degradation:

$$\begin{aligned}
\frac{d}{dt}[a_2.rC] &= \lambda^{+}_{a_2.rC}[a_2][rC] - \lambda^{-}_{a_2.rC}[a_2.rC] \\
\frac{d}{dt}[C] &= \lambda_{rC}[rC] + \lambda_{a_2.rC}[a_2.rC] + \lambda_{b_2.rC}[b_2.rC] - \lambda_C[C] \\
\frac{d}{dt}[c] &= \lambda_{Cc}[C] - \lambda_c[c], \qquad \frac{d}{dt}[c_2] = \lambda^{+}_{cc}[c]^2 - \lambda^{-}_{cc}[c_2]
\end{aligned} \tag{6.4}$$

Here, the λ_i are kinetic constants, available from the references in [155], t represents time, [.] means concentration, $a_2.rC$ and $b_2.rC$ represent transcription factors a_2 and b_2 bound to the *cis*-regulatory site rC, and the remaining symbols are explained in the caption of Figure 6.4. The system of differential equations (6.4) is taken from chemical kinetics ([3], Chapter 28). Consider, for instance, the formation and decay of a protein dimer: $c + c \leftrightarrow c_2$. The forward reaction

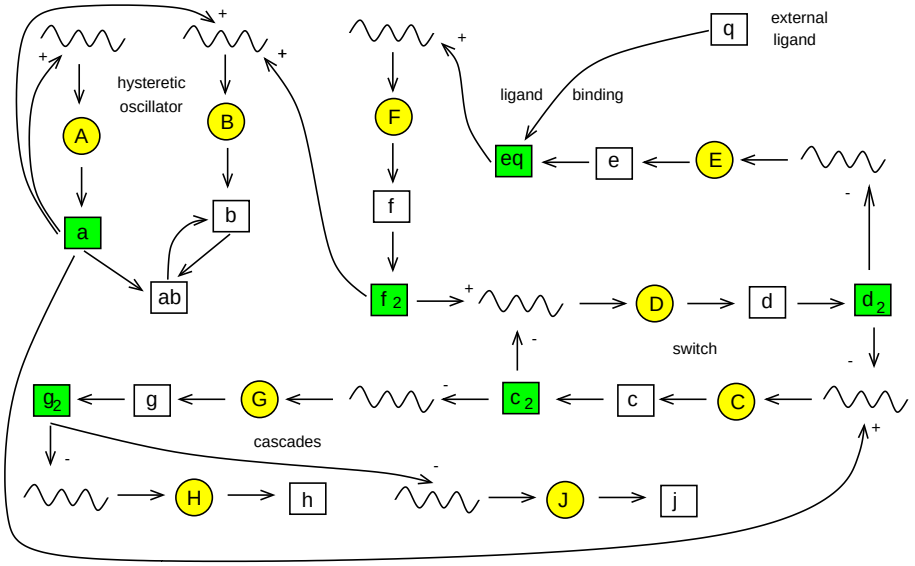


Figure 6.4: **Realistic biological network** composed of the elementary processes of Figure 6.3, taken from [155] in a slightly modified form. Oscillating lines represent *cis* regulatory sites in the gene upstream regions (promoters), mRNAs are symbolized by upper case letters in circles, proteins are shown by lower case letters in squares. Shaded squares indicate active transcription factors, where in all but two cases this activation is effected by dimerization, and in one case by ligand binding. The symbols + and − indicate whether a transcription factor acts as an activator or inhibitor. The network contains several subnetworks reported in the biological literature. The subnetwork involving mRNAs A and B is a hysteretic oscillator. A is translated into protein a, which is an active transcription factor that activates the transcription of B. B is translated into protein b, which forms a dimer ab. This dimerization reduces the amount of free transcription factors a, and oscillations result as a consequence of this negative feedback loop. The subnetwork involving mRNAs C and D is a switch: each mRNA is translated into a transcription factor that inhibits the transcription of the other mRNA, thereby switching the competing path "off". Finally, the subnetwork involving mRNA F is triggered by an external ligand, which is needed to form an active transcription factor dimer.

(formation) is second order, involving two monomers. Consequently, the time derivative of the dimer concentration, $\frac{d}{dt}[c_2]$, is proportional to the square of the concentration of the monomer, $[c]^2$. The reverse reaction (decay) is first order, and $\frac{d}{dt}[c_2]$ is proportional to the concentration of the dimer, $[c_2]$. Both processes together are described by the second equation in the last row of (6.4). The remaining equations can be explained similarly. The system of differential equations for the whole regulatory network of Figure 6.4 is composed of these elementary equations, with three additional but similar equations for ligand binding, ligand degradation, and heterodimerization ($a, b \leftrightarrow ab$). The resulting set of differential equations is stiff and needs to be integrated numerically with a high-order adaptable step-size method (e.g. Runge–Kutta–Fehlberg). Note that except for a, all transcription factors dimerize before they are active, that each gene has more than one rate of transcription, depending on whether promoters are bound or unbound, and that the presence of different time scales makes it representative of a real biological system and a suitable challenge for the Bayesian network inference algorithm. In contrast to the work described in [155], the system was augmented by adding 41 spurious, unconnected genes (giving a total of 50 genes), which were up- and down-regulated at random.

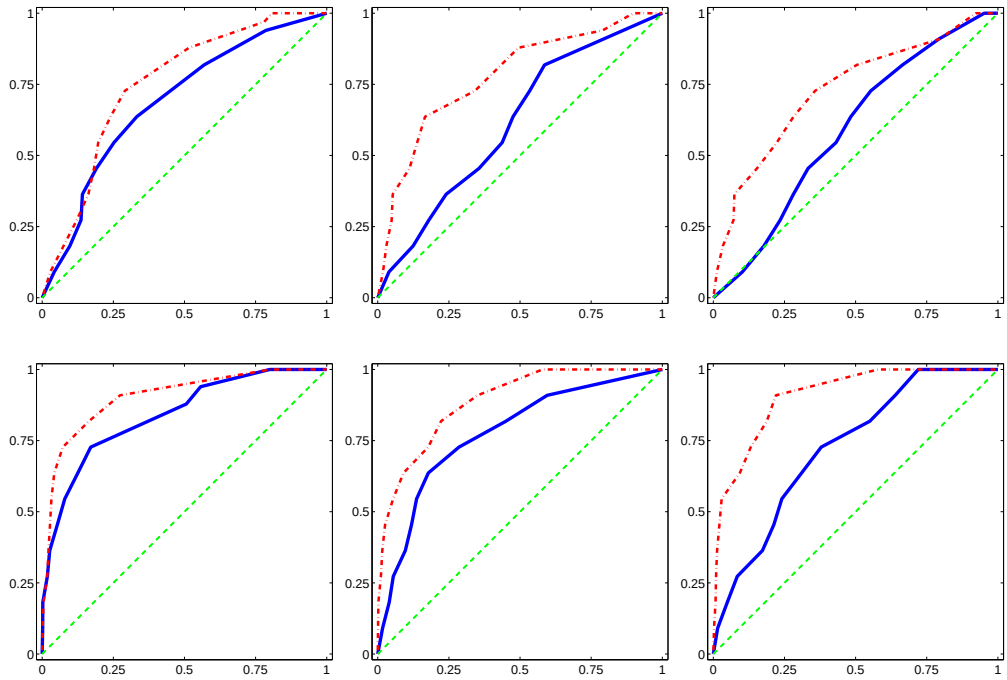


Figure 6.5: **ROC curves for the realistic simulated data,** averaged over three MCMC simulations. The rows represent different sampling periods. *Top row:* Sampling over a long time interval of 4000 minutes, which mainly covers the system in equilibrium. *Bottom row:* Restricting the sampling to a short 500-minute time interval immediately after ligand injection, when the system is in a perturbed non-equilibrium state. The columns correspond to different structure priors. *Left column:* maximum fan-in = 2; *middle column:* maximum fan-in = 3; *right column:* maximum fan-in = 4. In each subfigure the sensitivity (proportion of recovered true edges) is plotted against the complementary specificity (proportion of false edges). The thin, diagonal dashed line is the expected ROC curve of a random predictor. The solid line shows the ROC curve obtained from gene expression data alone, while the dash-dotted line shows the ROC curve obtained when including information about potential transcription factor binding motifs in gene upstream sequences. See [DH23] for more details.

The first experiment followed closely the procedure in [155]. Ligand was injected for 10 minutes at a rate of 10^5 molecules/minute at time 1000 minutes. Then, 12 data points were collected over 4000 minutes in equi-distant intervals. The second experiment adopted a sampling strategy different from [155]. An analysis of the mRNA abundance levels reveals regular oscillations when the system is in equilibrium. Such signals are known to have a low information content; consequently, it seems to make better sense to focus on the time immediately after external perturbation, when the system is in disequilibrium. The sampler therefore collected 12 data points over a shorter interval of only 500 minutes immediately after ligand injection, between times 1100 and 1600 minutes.

To learn the network structure, Bayesian networks $\mathcal{M}_i$ were sampled from the posterior distribution $P(\mathcal{M}|\mathcal{D})$ with MCMC. From a comparison between this sample, $\{\mathcal{M}_i\}$, and the known true network, one can estimate the accuracy of the inference procedure. Denote by $P(e_{ik}|\mathcal{D})$ the posterior probability of an edge e_{ik} between nodes i and k, which is given by the proportion of networks in

the MCMC sample $\{\mathcal{M}_i\}$ that contain this edge. Let $\mathcal{E}(\rho) = \{e_{ik}|P(e_{ik}|\mathcal{D}) > \rho\}$ denote the set of all edges whose posterior probability exceeds a given threshold $\rho \in [0, 1]$. From this set we can compute (1) the *sensitivity*, that is, the proportion of recovered true edges, and (2) the complementary *specificity*, that is, the proportion of erroneously recovered spurious edges. To rephrase this: For a given threshold ρ we count the number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) edges. We then compute the sensitivity $= TP/(TP + FN)$, the specificity $= TN/(TN+FP)$, and the complementary specificity = 1-specificity $= FP/(TN+FP)$. Rather than selecting an arbitrary value for the threshold ρ, we repeat this scoring procedure for several different values of $\rho \in [0, 1]$ and plot the ensuing sensitivity scores against the corresponding complementary specificity scores. This gives the *receiver operating characteristics* (ROC) curves of Figure 6.5. The diagonal dashed line indicates random expectation. A ROC curve following the vertical and horizontal axes from $[TP = 0, FP = 0]$ to $[TP = 1, FP = 0]$ and $[TP = 1, FP = 1]$ indicates perfect prediction. In general, ROC curves are between these two extremes, with a larger *area under the curve* (AUC) indicating a better performance.

The results are shown in Figure 6.5. A detailed analysis can be found in [DH23]. The main findings can be summarized as follows. The ROC curves are never perfect, which indicates that the true regulatory network cannot be learned in its entirety. On the other hand, the ROC curves are always better than random expectation, which suggests that certain local network features can be successfully learned, and that biological hypotheses can be assessed in a network-wide context. The network reconstruction accuracy depends on the sampling scheme. As conjectured above, sampling in disequilibrium after perturbation of the system by ligand injection leads to better ROC curves than sampling in equilibrium. The results also improve with a restriction of the number of potential regulators (fan-in restriction) and the inclusion of potential transcription factor binding motifs in the gene upstream sequences. The latter findings provide an example of how the network reconstruction accuracy can be boosted by the inclusion of biological prior knowledge. I will delve into this important topic in more detail in the next section.

In a second related study with Adriano Werhli and Marco Grzegorczyk [DH18], we carried out a comprehensive comparative evaluation of Bayesian networks, graphical Gaussian models (GGMs) [126, 127], and relevance networks [21, 22]. While Bayesian networks clearly outperformed relevance networks, in corroboration of Figure 6.1, we found that their performance was on a par with GGMs when data from passive observations were used. Only when active interventions in the form of gene knockouts or over-expressions were included, did Bayesian networks outperform GGMs. The details of this study are available from [DH18].

6.3 Bayesian Integration of Biological Prior Knowledge

The number of experimental conditions is usually limited for practical reasons. For instance, a biologist will hardly be able to procure the funding for, say, a thousand microarray experiments, and in the practical data analysis, we have to make do with a few dozen conditions. This sparsity of data will usually lead to a diffuse posterior distribution in network structure space, if the posterior distribution is dominated by the likelihood. A way to counteract this tendency is to include informative prior knowledge, as described in the Introduction section. For instance, a biologist aiming to infer regulatory networks in a crop plant, say barley, might want to include knowledge about related regulatory networks in a model plant like Arabidopsis. The bulk of the analysis may be based on gene expression profiles from microarray experiments, but we may have information about transcription factor binding motifs in the promoters and upstream sequences of some genes, which will give us prior cues about potential gene regulation patterns. It is certainly desirable to include this prior knowledge in the analysis. However, it is also important to automatically assess how useful this prior knowledge is. How confident are we that the over-representation of certain sequence motifs in the promoter is indicative of genuine transcription factor binding? To what extent is a signalling pathway in a model plant similar to corresponding pathways in crops, and is there any corresponding pathway in the first place? We would therefore like to have a mechanism in place that automatically trades off different sources of prior knowledge against each other and against the data, so as to assess their relative merits. This was the objective of my work with Adriano Werhli [DH13,DH16], where the aim was a systematic integration of biological prior knowledge in the inference of gene regulatory networks by following the Bayesian paradigm.

6.3.1 Method

A network $\mathcal{M}$ is represented by a binary adjacency matrix, where each entry $\mathcal{M}_{ij}$ can be either 0 or 1. A zero entry, $\mathcal{M}_{ij} = 0$, indicates the absence of an edge between node i and node j. Conversely if $\mathcal{M}_{ij} = 1$ there is a directed edge from node i to node j. We define the biological prior knowledge matrix B to be a matrix in which the entries $B_{ij} \in [0, 1]$ represent our knowledge about interactions between nodes as follows: If entry $B_{ij} = 0.5$, we do not have any prior knowledge about the presence or absence of the directed edge between node i and node j. If $0 \leq B_{ij} < 0.5$ we have prior evidence that the directed edge between node i and node j is absent. The evidence is stronger as B_{ij} is closer to 0. If $0.5 < B_{ij} \leq 1$ we have prior evidence that the directed edge pointing from node i to node j is present. The evidence is stronger as B_{ij} is closer to 1.

Following an approach proposed by Imoto et al. [68] we define a function that measures the

agreement between a given network $\mathcal{M}$ and our biological prior knowledge:

$$E(\mathcal{M}) = \sum_{i,j=1}^{N} |B_{i,j} - \mathcal{M}_{i,j}| \tag{6.5}$$

where N is the total number of nodes in the studied domain. Note that E is zero for a perfect match between the prior knowledge B and the actual network structure $\mathcal{M}$, while increasing values of E indicate an increasing mismatch between B and $\mathcal{M}$. We next define the prior distribution over network structures $\mathcal{M}$ to take the form of a Gibbs distribution:

$$P(\mathcal{M}|\beta) = \frac{e^{-\beta E(\mathcal{M})}}{Z(\beta)} \tag{6.6}$$

where $E(\mathcal{M})$ was defined in (6.5) and corresponds to an 'energy' in statistical physics, β is a hyperparameter that corresponds to an inverse temperature, and the denominator is a normalizing constant that is usually referred to as the partition function: $Z(\beta) = \sum_{\mathcal{M}\in\mathbb{M}} e^{-\beta E(\mathcal{M})}$. Note that this summation extends over the set of all possible network structures, $\mathbb{M}$. Since the number of structures increases superexponentially with the number of nodes (see e.g. Chapter 2 in [DH1]), the summation becomes soon intractable as N increases. For DBNs without intra-slice edges the computational costs can be reduced to polynomial time complexity by exploiting intrinsic modularities in the system, as shown in [DH16,DH42]. For static Bayesian networks the assumption of the same modularities leads to a tractable expression for $Z(\beta)$ that is an approximation and systematically over-estimates the true value; see [DH16] for details. The hyperparameter β can be interpreted as a factor that indicates the strength of the influence of the biological prior knowledge relative to the data. For $\beta \to 0$, the prior distribution defined in (6.6) becomes flat and uninformative about the network structure. Conversely, for $\beta \to \infty$, the prior distribution becomes sharply peaked at the network structure with the smallest deviation from B.

The approach can easily be generalized to multiple sources of prior knowledge. To keep the notation uncluttered, I restrict the exposition to two sources of prior knowledge; the extension to more than two sources is straightforward. Assume that the biological prior knowledge stems from two independent sources (e.g. TF binding motifs in the gene upstream sequences and pathways from KEGG). This can be represented by two separate prior knowledge matrices B^k, $k \in \{1, 2\}$, each satisfying the requirements laid out in the previous paragraphs. This gives us two 'energy' functions:

$$E_1(\mathcal{M}) = \sum_{i,j=1}^{N} \left|B^1_{i,j} - \mathcal{M}_{i,j}\right|; \qquad E_2(\mathcal{M}) = \sum_{i,j=1}^{N} \left|B^2_{i,j} - \mathcal{M}_{i,j}\right| \tag{6.7}$$

where each 'energy' is associated with its own hyperparameter β_k. The prior probability of a network $\mathcal{M}$ given the hyperparameters β_1 and β_2 is now defined as:

$$P(\mathcal{M}|\beta_1, \beta_2) = \frac{e^{-\{\beta_1 E_1(\mathcal{M}) + \beta_2 E_2(\mathcal{M})\}}}{Z(\beta_1, \beta_2)} \tag{6.8}$$

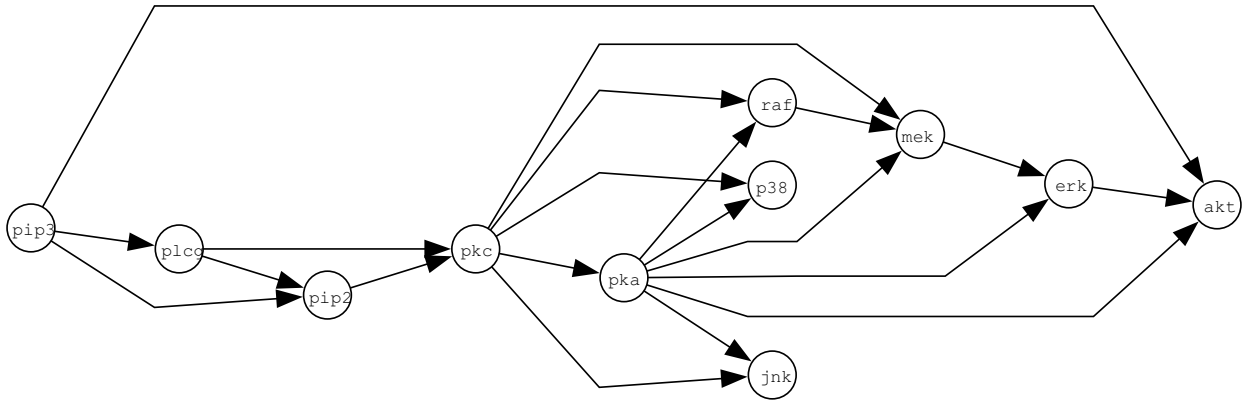


Figure 6.6: **Raf signalling pathway.** The graph shows the Raf signalling network, taken from [123]. Nodes represent proteins, edges represent interactions, and arrows indicate the direction of signal transduction.

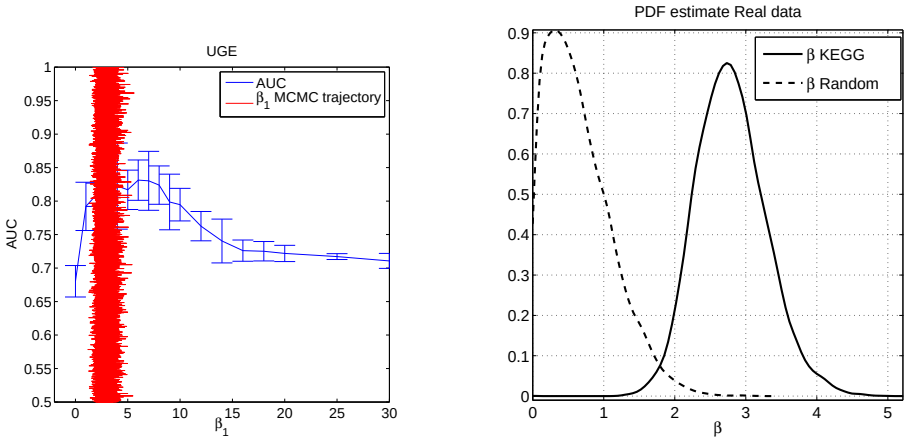
where the partition function is given by $Z(\beta_1, \beta_2) = \sum_{\mathcal{M} \in \mathbb{M}} e^{-\{\beta_1 E_1(\mathcal{M}) + \beta_2 E_2(\mathcal{M})\}}$. Given a set of transcriptional profiles $\mathcal{D}$ and our prior knowledge base $B = \{B^1, B^2\}$, the objective of Bayesian inference is to determine the posterior distribution $P(\mathcal{M}, \beta_1, \beta_2 | \mathcal{D}, B)$. This distribution is analytically intractable; so we resort to MCMC to sample network structures $\mathcal{M}$ and hyperparameters β_1, β_2 from it numerically. This approach makes use of the factorization

$$P(\mathcal{M}, \beta_1, \beta_2 | \mathcal{D}, B) \;\propto\; P(\mathcal{D}, \mathcal{M}, \beta_1, \beta_2, B) \;=\; P(\mathcal{D}|\mathcal{M}) P(\mathcal{M}|\beta_1, \beta_2, B) P(\beta_1, \beta_2 | B) \tag{6.9}$$

where $P(\beta_1, \beta_2 | B)$ denotes the prior on the hyperparameters which, in the absence of genuine prior knowledge, can be chosen uninformative. For full methodological and implementational details, see [DH16].

6.3.2 Empirical evaluation on the Raf signalling pathway

In [DH16], we carried out a comprehensive evaluation of the proposed method on a variety of synthetic and real data sets. As an illustration, I will present an application to the reconstruction of a protein signal transduction network. Sachs et al. [123] applied intracellular multicolour flow cytometry experiments to quantitatively measure protein concentrations. Data were collected after a series of stimulatory cues and inhibitory interventions targeting specific proteins in the Raf pathway. Raf is a critical signalling protein involved in regulating cellular proliferation in human immune system cells. The dysregulation of the Raf pathway is implicated in carcinogenesis, and this pathway has therefore been extensively studied in the literature [35, 123]; see Figure 6.6 for a representation of the gold-standard network from [123]. Note that the number of experimental replications for cytometry data substantially exceeds that of currently available gene expression data from microarrays. Our objective was to use the cytometry data as a proxy to assess the gene network reconstruction accuracy that one can expect to obtain from microarray experiments, drawing on the



(a) (b)

Figure 6.7: **Inferring hyperparameters from the cytometry data of the Raf pathway.** *Panel (a):* The horizontal axis represents the value of β_1, the hyperparameter associated with the prior knowledge from KEGG. The vertical axis represents the area under the ROC curve (AUC) for the undirected graph evaluation (UGE). The results for the directed graph evaluation (DGE) were similar. The graph plotted against the horizontal axis shows the mean AUC score for fixed values of β_1, obtained by sampling network structures from the posterior distribution with MCMC. The results were averaged over five data sets of 100 protein concentrations each, independently sampled from the observational cytometry data of Sachs et al. [123]. The error bars show the respective standard deviations. The graph plotted against the vertical axis shows trace plots of β_1 obtained by approximately sampling these values from the posterior distribution with MCMC. *Panel (b)* shows the corresponding posterior probability densities, estimated from the MCMC trajectories using a Parzen estimator with a Gaussian kernel. The solid line refers to the hyperparameter associated with the prior knowledge extracted from KEGG. The dashed line refers to random and hence vacuous prior knowledge. The data, on which the inference was based, consisted of 100 concentrations of the 11 proteins in the Raf pathway, subsampled from the observational cytometry data of Sachs et al. [123].

fact that we have a fairly reliable gold standard for the Raf network structure (see Figure 6.6). We therefore downsampled the data to a sample size representative of current microarray experiments (100 exemplars). In our experiments we used 5 data sets with 100 measurements each, obtained by randomly sampling subsets from the original observational data of Sachs et al. [123].[1]

We extracted biological prior knowledge from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways database [71, 72, 73]. KEGG pathways represent current knowledge of the molecular interaction and reaction networks related to metabolism, other cellular processes, and human diseases.[2] From these pathways, we computed the prior knowledge matrix, introduced in

[1] Details about the standardization of the data can be found [DH18].

[2] As KEGG contains different pathways for different diseases, molecular interactions and types of metabolism, it is possible to find the same pair of genes (I use the term "gene" generically for all interacting nodes in the network; this may include proteins encoded by the respective genes) in more than one pathway. We therefore extracted all

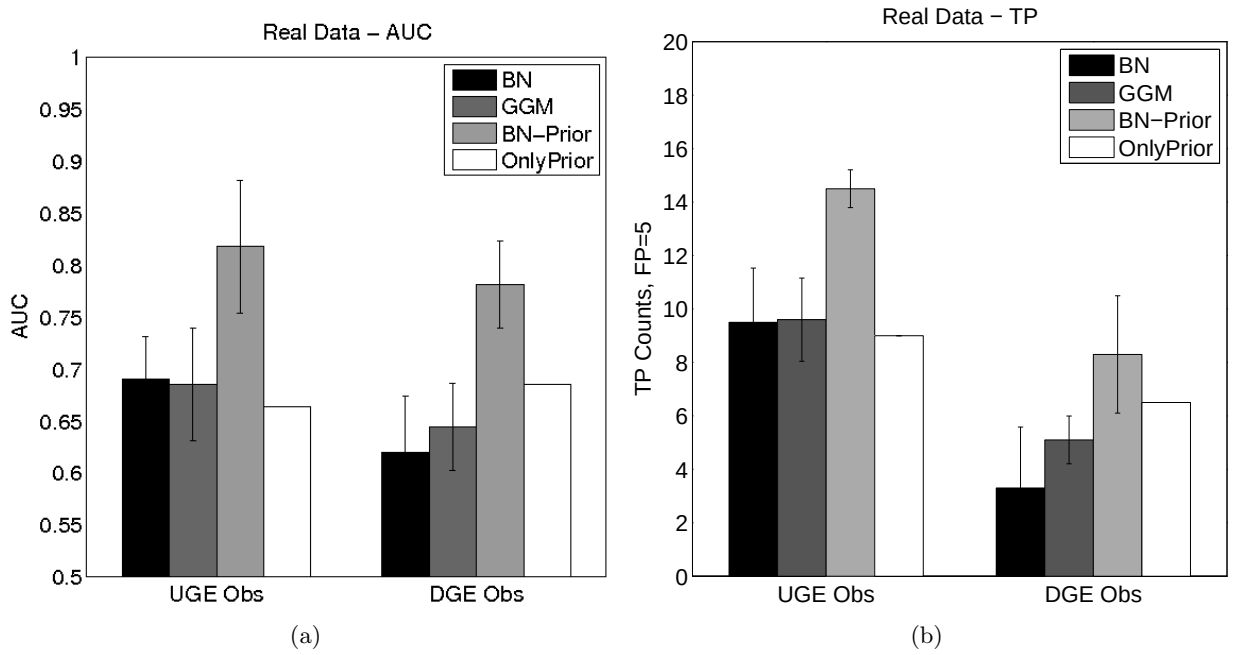


Figure 6.8: **Reconstruction of the Raf signalling pathway with different machine learning methods.** The figure evaluates the accuracy of inferring the Raf signalling pathway from cytometry data and prior information from KEGG. Two evaluation criteria were used. The left panel shows the results in terms of the area under the ROC curve (AUC scores), while the right panel shows the number of predicted true positive (TP) edges for a fixed number of 5 spurious edges. Each evaluation was carried out twice: with and without taking the edge direction into consideration (UGE: undirected graph evaluation, DGE: directed graph evaluation). Four machine learning methods were compared: Bayesian networks without prior knowledge ('BNs'), graphical Gaussian models without prior knowledge ('GGMs'), Bayesian networks with prior knowledge from KEGG ('BN-Prior'), and prior knowledge from KEGG only ('Only Prior'). In the latter case, the elements of the prior knowledge matrix were computed from equation (6.10). The histogram bars represent the mean values obtained by averaging the results over five data sets of 100 protein concentrations each, independently sampled from the observational cytometry data of Sachs et al. [123]. The error bars show the respective standard deviations.

Subsection 6.3.1, as follows. Define by U_{ij} the total number of times a pair of genes i and j appears in a pathway, and by u_{ij} the number of times the genes are connected by a (directed) edge in the KEGG pathway. The elements B_{ij} of the prior knowledge matrix are then defined by

$$B_{ij} \;=\; \frac{u_{ij}}{U_{ij}} \tag{6.10}$$

If a pair of genes is not found in any of the KEGG pathways, we set the respective prior association to $B_{ij} = 0.5$, implying that we have no prior information about this relationship.

The objective of the empirical evaluation in [DH16] was to assess the viability of the Bayesian inference scheme and to estimate by how much the network reconstruction accuracy improves as a consequence of combining the (down-sampled) cytometry data with prior knowledge from the KEGG pathway database. To this end, we compared the results obtained with the proposed Bayesian model for the integration of biological prior knowledge, with Bayesian networks (BNs) and graphical Gaussian models (GGMs) that were trained without the inclusion of prior knowledge. The latter were applied as described in [127]. Full details of the methodologies, simulations and comparative evaluation are available from [DH16].

There are three interesting questions that we wanted to answer: (1) Can the proposed approach discriminate between useful and vacuous prior knowledge? (2) Do we obtain the optimal trade-off between prior knowledge and data? (3) Do we get an improvement in the network reconstruction accuracy?

Question 1: Can the proposed approach discriminate between useful and vacuous prior knowledge?

We wanted to test whether the proposed Bayesian inference method can discriminate between different sources of prior knowledge and automatically assess their relative merits. To this end, we complemented the prior knowledge from the KEGG pathway database with a second source of prior knowledge, for which the entries in the prior knowledge matrix were chosen at random. Hence, this second source of prior knowledge is vacuous and does not include any useful information for reconstructing the regulatory network. Figure 6.7(b) shows the estimated posterior distributions of two hyperparameters: β_1, the hyperparameter associated with the prior knowledge extracted from KEGG; and β_2, the hyperparameter associated with the vacuous source of prior knowledge. It is seen that the hyperparameter associated with KEGG, β_1, takes on substantially larger values than the hyperparameter associated with the vacuous prior information, β_2. This suggests that the proposed method successfully discriminates between the two sources of prior information and effectively suppresses the influence of the vacuous one.

pathways from KEGG that contained at least one pair of the 11 proteins/phospholipids included in the Raf pathway. We found 20 pathways that satisfied this condition.

Question 2: Do we obtain the optimal trade-off between prior knowledge and data?

While the study described in the previous paragraph suggests that the proposed Bayesian inference scheme succeeds in suppressing irrelevant prior knowledge, we were curious to see whether the hyperparameter associated with the relevant source of prior knowledge (from KEGG) was optimally inferred. To this end, we chose a large set of fixed values for β_1, while keeping the hyperparameter associated with the vacuous prior information fixed at zero: $\beta_2 = 0$. For each fixed value of β_1, we sampled Bayesian network structures from the posterior distribution with MCMC, and evaluated the network reconstruction accuracy using the evaluation criteria described below. We compared these results with the proposed Bayesian inference scheme, where both hyperparameters and networks are simultaneously sampled from the posterior distribution with MCMC. The results are shown in Figure 6.7(a). They suggest that the inferred values of β_1 are close to those that achieve the best network reconstruction accuracy, and that the approximation of the partition function, as briefly described in Section 6.3.1, leads only to a small bias.

Question 3: Do we get an improvement in the network reconstruction accuracy?

While the true network is a directed graph, our reconstruction methods may lead to undirected, directed, or partially directed graphs[3]. To assess the performance of these methods, we applied two different criteria. The first approach, referred to as the *undirected graph evaluation* (UGE), discards the information about the edge directions altogether. To this end, the original and learned networks are replaced by their skeletons, where the skeleton is defined as the network in which two nodes are connected by an undirected edge whenever they are connected by any type of edge. The second approach, referred to as the *directed graph evaluation* (DGE), compares the predicted network with the original directed graph. A predicted undirected edge is interpreted as the superposition of two directed edges, pointing in opposite directions. The application of any of the machine learning methods considered in our study leads to a matrix of scores associated with the edges in a network. For Bayesian networks sampled from the posterior distribution with MCMC, these scores are the marginal posterior probabilities of the edges. For GGMs, these are partial correlation coefficients. Both scores define a ranking of the edges. This ranking defines a receiver operating characteristics (ROC) curve, where the relative number of true positive (TP) edges is plotted against the relative number of false positive (FP) edges; see Section 6.2 for details. We proceeded with two different evaluation procedures. The first approach is based on integrating the ROC curve so as to obtain the area under the curve (AUC), with larger areas indicating, overall, a better performance. The second approach is based on the selection of an arbitrary threshold on the edge scores, from which a specific network prediction is obtained. In [DH16], we set the threshold such that it led to a fixed

[3]GGMs are undirected graphs. While Bayesian networks are, in principle, directed graphs, partially directed graphs may result as a consequence of equivalence classes, which were briefly mentioned in Section 6.1. For more details, see e.g. Chapter 2 in [DH1].

count of 5 FPs. From the predicted network, we determined the number of correctly predicted (TP) edges, and took this score as our second figure of merit.

The results are shown in Figure 6.8. The proposed Bayesian inference scheme clearly outperforms the methods that do not include the prior knowledge from the KEGG database (Bayesian networks and GGMs). It also clearly outperforms the prediction that is solely based on the KEGG pathways alone without taking account of the cytometry data. The improvement is significant for all four evaluation criteria: AUC and TP scores for both directed (DGE) and undirected (UGE) graph evaluations. This suggests that the network reconstruction accuracy can be substantially improved by systematically integrating expression data with prior knowledge about pathways, as extracted from the literature or databases like KEGG.

Conclusion

Our simulation study gave a clear affirmative response to all three questions, suggesting that the proposed method adds an important tool to the kit of computational systems biology.

6.4 Active Pathways under Different Experimental Conditions

The assumption so far has been that the molecular biological system of interest can be characterized by a unique regulatory network. What we are actually aiming to infer, though, are the active parts of this network, which may differ under different experimental conditions. To illustrate this point, consider a transcription factor that potentially upregulates a group of genes further downstream in the regulatory chain. If the experimental conditions are chosen such that the gene coding for this transcription factor is never expressed itself, then the respective subnetwork will never be activated, and hence cannot be inferred from the data. When aiming to infer regulatory networks related to an organism's immune system, we would expect certain pathways to be activated only upon infection, and remaining invisible when gene expression profiles are only taken in the healthy state. In fact, some preliminary analysis in [150] related to the challenging of macrophages with interferon gamma (IFNγ) and viral infection has revealed differences in the active pathways under the conditions of viral infection, IFNγ treatment, and viral infection plus IFNγ treatment. This suggests that a regulatory network is not an immutable entity, but may vary in response to changes in the experimental and/or environmental conditions.

When aiming to reconstruct a network from gene expression profiles taken under different experimental conditions, there seem to be two principled approaches we may pursue. The first is to ignore

the changes in the experimental conditions altogether and merge the data into one monolithic set. The problem with this approach is that it inevitably blurs the differences between the different conditions and thereby obscures the biological insight we are aiming to gain; for instance, we would not be able to tell the difference between the state of a network in infected, healthy, and IFNγ-treated cells. The second approach is to keep the data obtained under different conditions separate, and to infer separate regulatory networks active under these different conditions. While this approach has the potential to reveal the differences between the regulatory networks in different states, e.g. infection versus treatment, it will almost inevitably result in a considerable reduction in statistical power and reconstruction accuracy. Current postgenomic data sets are usually sparse, e.g. the number of microarray experiments biologists can afford to carry out is usually limited to the order of a few dozen, which compromises the extent to which networks can be reconstructed. Breaking a sparse data set up into smaller units will inevitably aggravate this situation and increase the uncertainty about inferred network structures.

In my work with Adriano Werhli described in [DH13], we aimed to pursue a compromise between the two extreme procedures described above. The motivation was given by the insight gained from an earlier study described in Chapter 2 in [150]. Although we found differences between the active pathways under the different conditions of infection and treatment with IFNγ, the networks shared considerable features they had in common. Our conjecture was that this holds in general, and that a cell's regulatory networks, while potentially transitioning between different active states in response to different external cues, share substantial features owing to a common generic network architecture. Our objective was to formulate this proposition mathematically so as to integrate it into the probabilistic modelling process.

As it turns out, this objective can be achieved by a modification of the probabilistic model described in the previous section, Section 6.3. Recall that the objective of Section 6.3 was the integration of explicit prior knowledge into the inference scheme by softly constraining the inferred network to be similar to the *a priori* known network. In modification of this scheme, we propose in [DH13] learning separate regulatory networks from disjunct gene expression data, but tying these networks together by softly constraining them to be similar to a shared underlying generic network. This approach overcomes the rigidity of the first scenario described above, which would obscure the differences between the network states in different experimental conditions. By sharing information between the different network states, the problem of the second scenario described above should be averted, that is, the statistical power and accuracy of the reconstruction should be considerably enhanced.

6.4.1 Method

In order to integrate information from I different data sets $(\mathcal{D}_1 \dots \mathcal{D}_I)$ obtained under different experimental conditions we use the probabilistic graphical model presented in Figure 6.9. Each

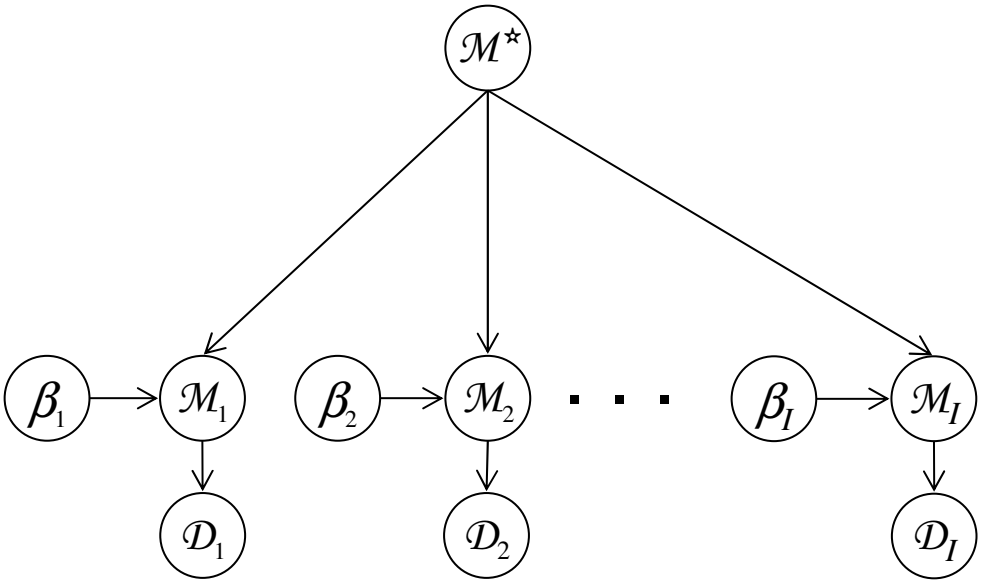


Figure 6.9: **Probabilistic graphical model for learning active subnetworks under different experimental conditions.** $(\mathcal{D}_1 \dots \mathcal{D}_I)$ are data sets obtained under different experimental conditions. Each of these data sets is associated with its own hyperparameter $(\beta_1, \dots, \beta_I)$ and network structure $(\mathcal{M}_1, \dots, \mathcal{M}_I)$. The hypernetwork $\mathcal{M}^\star$ leads to a coupling between the individual network structures $(\mathcal{M}_1, \dots, \mathcal{M}_I)$ and encourages them to be similar.

data set $(\mathcal{D}_1 \dots \mathcal{D}_I)$ is associated with its own hyperparameter $(\beta_1, \dots, \beta_I)$ and network structure $(\mathcal{M}_1, \dots, \mathcal{M}_I)$. The latent graph $\mathcal{M}^\star$, which is not directly associated with the data, leads to a coupling between the individual network structures $(\mathcal{M}_1, \dots, \mathcal{M}_I)$ and encourages them to be similar. Note that Figure 6.9 constitutes a hierarchical Bayesian model, in which the β_is and $\mathcal{M}^\star$ correspond to hyperparameters that determine the prior distribution on the network structures $\mathcal{M}_i$s. Further note that $\mathcal{M}^\star$ is not just a variable, but a complex entity representing a whole network itself. I shall therefore refer to $\mathcal{M}^\star$ as the hypernetwork.

The joint probability factorizes according to the independence relations defined by the probabilistic graphical model of Figure 6.9 as follows:

$$P(\mathcal{M}_1, \dots, \mathcal{M}_I, \mathcal{D}_1 \dots \mathcal{D}_I, \beta_1, \dots, \beta_I, \mathcal{M}^\star) \;=\; P(\mathcal{M}^\star) \prod_{i=1}^{I} P(\mathcal{D}_i|\mathcal{M}_i) P(\mathcal{M}_i|\beta_i, \mathcal{M}^\star) P(\beta_i) \quad (6.11)$$

where the prior distribution over network structures, $P(\mathcal{M}_i|\beta_i, \mathcal{M}^\star)$, takes the form of a Gibbs distribution:

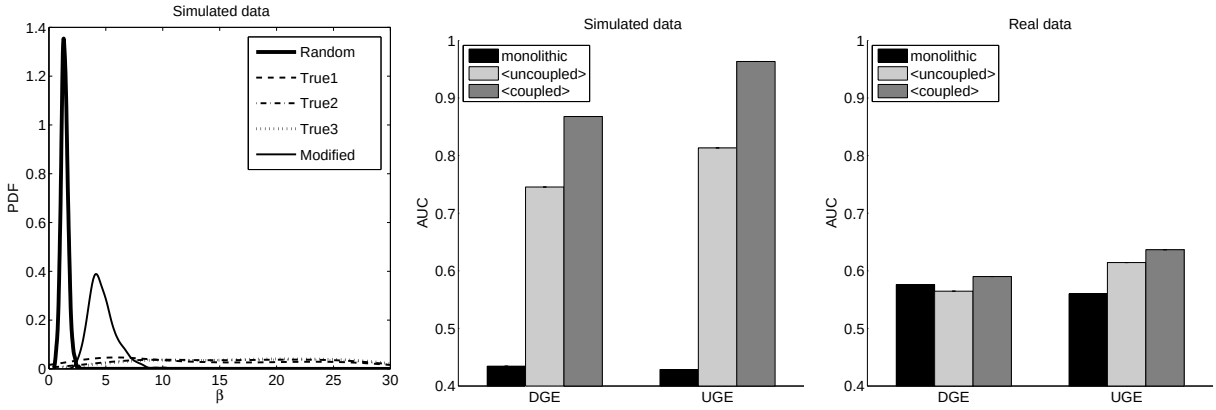
$$P(\mathcal{M}_i|\beta_i, \mathcal{M}^\star) = \frac{e^{-\beta_i(|\mathcal{M}_i - \mathcal{M}^\star|)}}{Z(\beta_i, \mathcal{M}^\star)}. \quad (6.12)$$

The term $|\mathcal{M}_i - \mathcal{M}^\star|$ measures the similarity between the graphs $\mathcal{M}_i$ and $\mathcal{M}^\star$, for instance in terms of the Hamming distance, i.e. the number of different edges. This scheme introduces a cou-

pling between the individual networks $\mathcal{M}_i$: deviations between $\mathcal{M}_i$ and $\mathcal{M}^\star$ are penalized, which implies an indirect penalty for deviations between $\mathcal{M}_i$ and $\mathcal{M}_k$, $i \neq k$. The hyperparameter β_i can be interpreted as a factor that indicates the strength of the influence of the hypernetwork $\mathcal{M}^\star$ relative to the data. For $\beta_i \to 0$, the prior distribution defined in equation (6.12) becomes flat and uninformative about the network structure. Conversely, for $\beta_i \to \infty$, the prior distribution becomes sharply peaked, forcing the network structure $\mathcal{M}_i$ to be equal to the hypernetwork $\mathcal{M}^\star$. The denominator in equation (6.12) is a normalizing constant, also known as the partition function: $Z(\beta_i, \mathcal{M}^\star) = \sum_{\mathcal{M}_i \in \mathbb{M}} e^{-\beta_i(|\mathcal{M}_i - \mathcal{M}^\star|)}$, where $\mathbb{M}$ is the set of all valid network structures. Its computation was already briefly discussed in the previous section, Subsection 6.3.1. For a comprehensive mathematical treatment, see [DH13,DH16]. The objective of Bayesian inference is to infer the posterior distribution of the networks associated with the different structures, the hypernetwork and the hyperparameters, $P(\mathcal{M}_1, \ldots, \mathcal{M}_I, \beta_1, \ldots, \beta_I, \mathcal{M}^\star | \mathcal{D}_1 \ldots \mathcal{D}_I)$. This is analytically intractable; so we exploit the factorization inherent in (6.11) and set up an MCMC simulation to approximately generate a sample from the posterior distribution. The mathematical details are presented in [DH13].

6.4.2 Findings

To simulate active pathways under different experimental conditions, we combined five individual data sets as follows. Three of the data sets were generated from the gold-standard RAF regulatory network, shown in Figure 6.6. A fourth data set was generated from a slightly modified version of this network, in which the following four edges had been deleted: PKC → RAF, PKC → PKA, PKA → MEK, and PLCg → PIP2. The deletion of these edges corresponds to changes in the active subpathways under different external conditions, as described above. As a fifth data set, we included a purely random data set. This corresponds to either a drastic change of the external conditions that deactivates the whole pathway, or to a flawed experiment that has corrupted the data. We wanted to investigate whether the proposed method succeeds in identifying this outlying data set and prevents it from adversely affecting the overall inference. We were also interested in whether the proposed method can distinguish between the data from the gold-standard and the modified RAF regulatory network. A description of the generation of the synthetic data is given in [DH13]; each subset contained 100 exemplars. For the cytometry data, we took four subsets of unintervened data, randomly selected from the data in [123] and pre-processed as described in [DH18]. Each subset contained 20 measurements. To these data sets we added a fifth data set of equal size, consisting of pure noise. The results are shown in Figure 6.10. The random data set deliberately included with the proper data is clearly detected. The hyperparameter associated with the random data is automatically set to very small values; this suggests that the proposed Bayesian coupling scheme is effective in switching off the influence of corrupted data. A data set generated from the modified network structure is also automatically detected. The associated hyperparameter is sampled from a distribution placed between those associated with the random data and the data generated from the unmodified network, successfully distinguishing it from both.



(a) (b) (c)

Figure 6.10: **Integrating different experimental conditions.** *Panel (a)* shows the posterior distributions of the hyperparameters $\beta_1, \ldots, \beta_5$ for the simulated data, estimated with a kernel estimator applied to the samples obtained from the MCMC simulations. The distribution of β_1, which is associated with the random data, is centred on small values close to zero. The hyperparameters β_2, β_3 and $\beta 4$, which are associated with the data sets generated from the true network, have a broad distribution reaching up to very large values. The distribution of β_5, which is associated with the modified network, is expected to lie between these two distributions, and this is in fact borne out in our simulations. The histograms in *Panels (b-c)* show a comparison of the network reconstruction accuracy in terms of the areas under the ROC curves (AUC scores) for three different methods: the monolithic approach (black), the uncoupled approach (light grey), and the proposed Bayesian coupling scheme (dark grey); see the main text for further details. The two panels correspond to different data sets: simulated data (Panel (b)), with a known gold standard, and protein concentrations from cytometry experiments (Panel (c)), with the network in Figure 6.6 taken as the gold standard. Each panel contains two histograms, evaluating only the reconstruction of the skeleton of the graph (UGE score) and additionally taking the edge direction into account (DGE score).

In terms of network reconstruction accuracy, the proposed Bayesian coupling scheme consistently outperforms the two competing approaches. The performance difference is clearly evident on the simulated data, where the individual data sets correspond to different activation levels of the regulatory subpathways (owing to different settings of the interaction parameters). But even for the cytometry data, where we only added one corrupted data set to four data sets obtained under homogeneous experimental conditions, the improvement is noticeable.

6.5 Non-Stationary Processes and Time-Varying Networks

In the previous section I have discussed the inference of network structures associated with different experimental or exogenous conditions. In the present chapter I will describe methods for

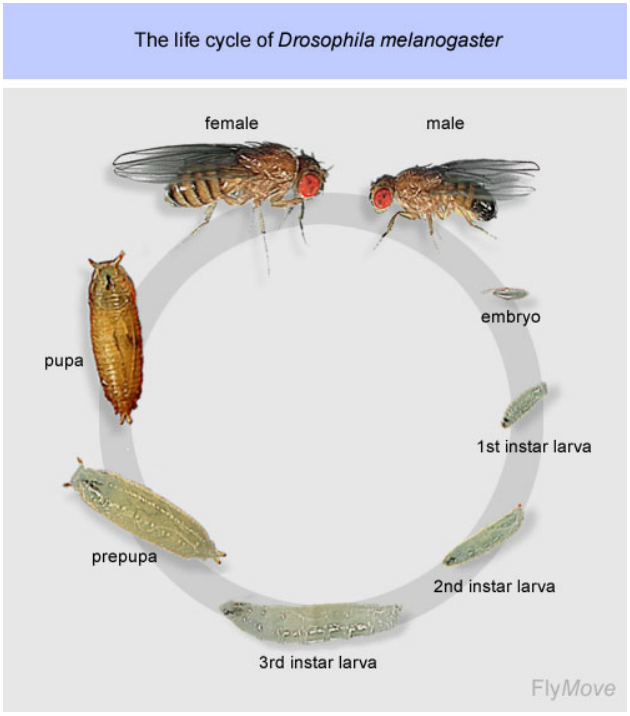
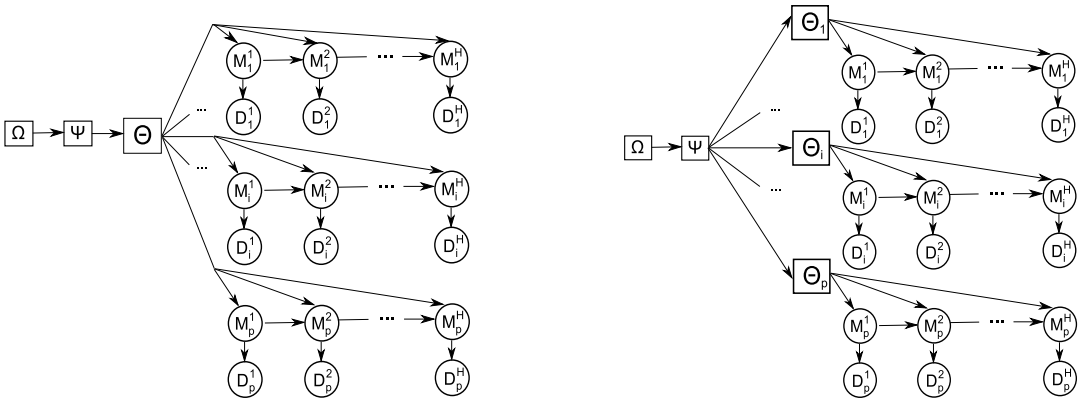


Figure 6.11: **Life cycle of Drosophila.** The figure shows the principal morphogenic stages during the life cycle of *Drosophila melanogaster*. Image taken from http://www.hoxfulmonsters.com.

inferring time-varying network structures during an evolutionary process. Consider the life cycle of *Drosophila melanogaster*, shown in Figure 6.11. It is certainly suboptimal to assume that the gene regulatory networks related to, say, wing muscle control are fixed and immutable during the organism's life cycle. What we need is a method that allows for changes in the regulatory network structure. Given that time series of high-throughput gene expression profiles are typically sparse, we want to prevent overfitting with a regularization scheme that incorporates our prior knowledge about the evolutionary nature of the process. Stated differently, we want a method that penalizes differences between network structures that are associated with adjacent time series segments. We also would like to generalize the method of the previous section and learn the number and location of the changepoints automatically from the data in an unsupervised manner. This was the objective of my work in [DH41,DH42], which I shall summarize in the present section.

6.5.1 Method

The standard assumption underlying dynamic Bayesian networks (DBNs), depicted in Figure 6.2, is that time-series have been generated from a homogeneous Markov process. To address this



(a) Hard Node Coupling (b) Soft Node Coupling

Figure 6.12: **Hierarchical Bayesian models for inter-segment and inter-node information coupling.** *Left panel*: Hard coupling between nodes, with a common hyperparameter Θ regulating the strength of the coupling between structures associated with adjacent segments, $\mathcal{M}_i^h$ and $\mathcal{M}_i^{h+1}$. *Right panel*: Soft coupling between nodes, with node-specific hyperparameters Θ_i coupled via level2-hyperparameters Ψ. The prior distribution of Ψ depends itself on some higher-level hyperparameters Ω. The changepoints are node-specific and divide the gene expression time series into segments. D_i^h denotes the data pertinent to node i in segment h. Full mathematical details of the model are presented in [DH41].

limitation, Lèbre et al. [83, 84] introduced a Bayesian multiple changepoint process, as described in Chapter 4 and illustrated in Figure 4.4, by which gene expression time series are divided into segments in an unsupervised manner. Different time series segments are treated as independent and are associated with separate network structures and parameters. A truncated Poisson prior is imposed on the number of changepoints. An RJMCMC [54] scheme based on changepoint birth and death moves is then applied to sample the number of changepoints, their location and the network structures associated with the different segments from the posterior distribution. The high flexibility of this approach can cause problems when applied to short time series with low numbers of measurements, as typically available from high-throughput experiments in systems biology, because it can lead to overfitting or inflated inference uncertainty.

In my joint work with Frank Dondelinger and Sophie Lèbre [DH41,DH42], I have addressed this shortcoming by introducing the concept of information sharing from the previous section. Due to the different nature of the data – a non-stationary process evolving in time – the central coupling depicted in Figure 6.9 is superseded by sequential coupling. Rather than encouraging all network structures to be similar to a common unknown hypernetwork, as in Figure 6.9, similarity is encouraged to hold for two network structures associated with adjacent time series segments. An illustration is given in Figure 6.12. We modified the prior from the previous section to allow for the fact that we are dealing with dynamic rather than static Bayesian networks. We compared two functional forms of the prior – an exponential versus a binomial distribution – and two coupling schemes: hard versus soft coupling, as illustrated in Figure 6.12. The coupling strengths depend

on various hyperparameters, which are organized via a hierarchical Bayesian model. Inference was carried out following the Bayesian paradigm, sampling changepoints, network structures and hyperparameters from the posterior distribution with RJMCMC. Full methodological details are given in [DH41,DH42].

6.5.2 Findings

Morphogenesis in *Drosophila melanogaster*

We applied our methods to a gene expression time series for eleven genes involved in the muscle development of *Drosophila melanogaster* [2]. The microarray data measured gene expression levels during all four major stages of morphogenesis: embryo, larva, pupa and adult. We investigated whether our methods were able to infer the correct changepoints corresponding to the known transitions between stages. Figure 6.13(a) shows the posterior probabilities of inferred changepoints obtained with a non-homogeneous DBN without information sharing [83, 84], while Figure 6.13(c) shows the posterior probabilities inferred with the information sharing methods. For comparison, we applied TESLA, a method based on L1-penalized regression proposed in [1] (Figure 6.13(b)). Our non-homogeneous DBNs are generally more successful than TESLA, in that they recover changepoints for all three transitions (embryo → larva, larva → pupa, and pupa → adult). Figure 6.13(b) indicates that the last transition, pupa → adult, is less clearly detected with TESLA, and it is missing in the related work of Robinson and Hartemink [118]. Both our method as well as TESLA detect additional transitions during the embryo stage, which are missing in [118]. We would argue that a complex gene regulatory network is unlikely to transition into a new morphogenic phase all at once, and some pathways might have to undergo activational changes earlier in preparation for the morphogenic transition. As such, it is not implausible that additional transitions at the gene regulatory network level occur. However, a failure to detect known morphogenic transitions can clearly be seen as a shortcoming of a method, and on these grounds our model appears to outperform the two alternative ones. We note that the main effect of information sharing is to reduce the size of the smaller peaks, while keeping the three most salient peaks (corresponding to larva → pupa, and pupa → adult, and an extra transition in the embryo phase). This reflects the fact that these changepoints are associated with significant changes in network structure, and adds to the interpretability of the results. The drawback is that the third morphological transition (embryo → larva) is less pronounced.

Reconstruction of a synthetic gene regulatory network in *Saccharomyces cerevisiae*

The highly topical field of synthetic biology enables biologists to design known gene regulatory networks in living cells. In the work described in [24], a synthetic regulatory network of 5 genes was constructed in *Saccharomyces cerevisiae* (yeast), and gene expression time series were measured

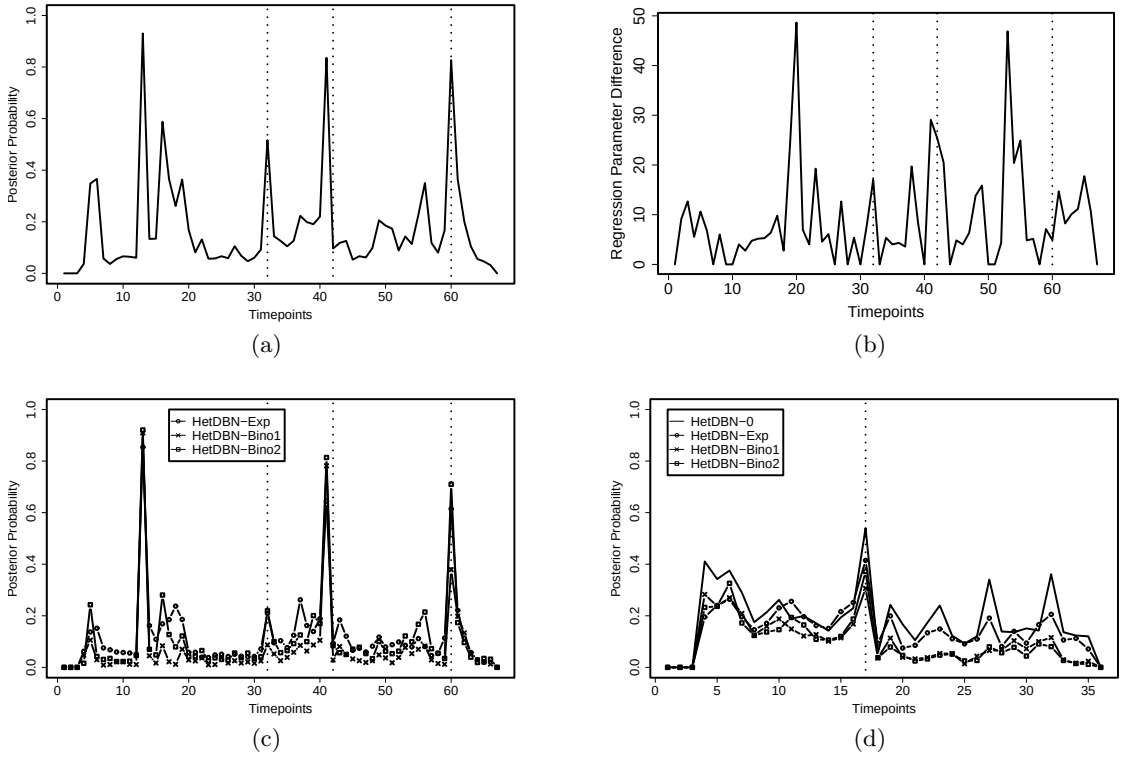


Figure 6.13: **Changepoints inferred on gene expression data related to morphogenesis in *Drosophila melanogaster*, and synthetic biology in *Saccharomyces cerevisiae* (yeast).** All figures obtained from DBNs show the posterior probability of a changepoint occurring for any node at a given time plotted against time. *Panel a:* Changepoints for Drosophila, obtained with a non-homogeneous DBN without information sharing, as proposed in [83, 84]. *Panel b:* TESLA, L1-norm of the difference of the regression parameter vectors associated with two adjacent time points plotted against time. *Panel c:* Changepoints for Drosophila inferred using the non-homogeneous DBNs with information sharing, as proposed in [DH41]. *Panel d:* Changepoints for the synthetic yeast strain, inferred using non-homogeneous DBNs. Four schemes are compared. HetDBN-0: no coupling. HetDBN-Exp: hard coupling based on an exponential prior. HetDBN-Bino1: hard coupling based on a binomial prior. HetDBN-Bino2: soft coupling based on a binomial prior. The notion of hard versus soft coupling is illustrated in Figure 6.12. See [DH41] for further methodological details. In Panels a-c, the vertical dotted lines indicate the three morphogenic transitions, while in Panel d the line indicates the boundary between "switch on" and "switch off" data.

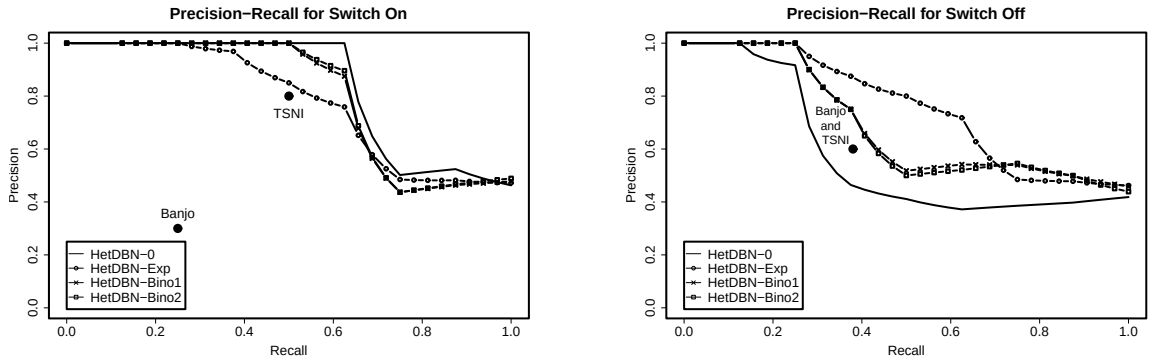


Figure 6.14: **Reconstruction of a known gene regulatory network from synthetic biology in *Saccharomyces cerevisiae*.** The network was reconstructed from two gene expression time series obtained with RT-PCR in two experimental conditions, reflecting the switch in the carbon source from galactose ("switch on") to glucose ("switch off"). For details, see [24]. Different information sharing and network coupling schemes, proposed in [DH41], were compared. HetDBN-0: no coupling. HetDBN-Exp: hard coupling based on an exponential prior. HetDBN-Bino1: hard coupling based on a binomial prior. HetDBN-Bino2: soft coupling based on a binomial prior. The notion of hard versus soft coupling is illustrated in Figure 6.12. For further methodological details, see [DH41]. The reconstruction accuracy is shown in terms of precision (vertical axis) - recall (horizontal axis) curves. Results were averaged over 10 independent MCMC simulations. The average areas under the PR curves, averaged over both phases ("switch on and off"), are as follows. HetDBN-0= 0.70, HetDBN-Exp= 0.77, HetDBN-Bino1= 0.75, HetDBN-Bino2= 0.75. For comparison, fixed precision/recall scores are shown for two state-of-the-art methods reported in [24]: Banjo, a conventional DBN, and TSNI, a method based ordinary differential equations (ODEs).

with RT-PCR for 16 and 21 time points under two experimental conditions, related to the carbon source: galactose ("switch on") and glucose ("switch off"). The authors tried to reconstruct the known gold-standard network from these time series with two established state-of-the-art methods from computational systems biology, one based on ordinary differential equations (ODEs), called TSNI, the other based on conventional DBNs, called Banjo; see [24] for details. Both methods are optimization-based and output a single network. By comparison with the known gold standard, the authors obtained the precision (proportion of predicted interactions that are correct) and recall (proportion of predicted true interactions) scores. In our study, we merged the time series from the two experimental conditions under exclusion of the boundary point[4], and applied the four non-homogeneous DBNs described before. Figure 6.13(d) shows the inferred marginal posterior probability of potential changepoints. The most significant changepoint is at the boundary between "switch on" and "switch off" data, confirming that the known true changepoint is consistently identified. The biological mechanism behind the other peaks is not known, and they are potentially spurious. Interestingly, the application of the proposed information-coupling schemes reduces the height of these peaks, with the binomial models having a stronger effect than the exponential one.

As we pursue a Bayesian inference scheme, we also obtain a ranking of the potential gene interactions in terms of their marginal posterior probabilities. From this ranking we computed the precision-

[4]When merging two time series $(x_1, \ldots, x_m)$ and $(y_1, \ldots, y_n)$, only the pairs $x_i \to x_j$ and $y_i \to y_j$ are presented to the DBN, while the pair $x_m \to y_1$ is excluded due to the obvious discontinuity.

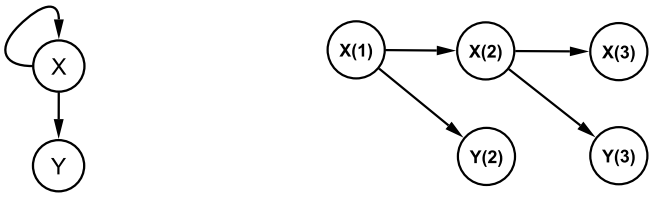


Figure 6.15: **State space graph and corresponding dynamic Bayesian network.** Left: Recurrent state space graph containing two nodes. Node X has a recurrent feedback loop and acts as a regulator of node Y. Right: The same graph unfolded in time. Compare with Figure 6.2.

recall (PR) curves [31] shown in Figure 6.14. Note that PR curves have recently become popular as an alternative to ROC curves, which I have discussed in Section 6.2, as they have certain methodological advantages; see the discussion in [31]. Our study shows that non-homogeneous DBNs with information sharing outperform Banjo and TSNI both in the "switch on" and the "switch off" phase. They also perform better than HetDBN-0 on the "switch off" data, but are slightly worse on the "switch on" data. Note that the reconstruction accuracy on the "switch off" data is generally poorer than on the "switch on" data [24]. Our results are thus plausible, suggesting that information sharing boosts the reconstruction accuracy on the poorer time series segment at the cost of a degraded performance on the stronger one. This effect is more pronounced for the exponential prior than for the binomial one, indicating a tighter coupling. The average areas under the PR curves, averaged over both phases ("switch on and off"), are shown in the caption of Figure 6.14. They show that the overall effect of information sharing is a performance improvement.

6.6 Modelling Nonlinear Regulation

Feedback loops and recurrent structures are essential to the regulation and stable control of complex biological systems. As discussed in Section 6.1 and illustrated in Figure 6.2, the application of dynamic as opposed to static Bayesian networks is promising in that, in principle, these feedback loops can be learned. However, in my joint work with Marco Grzegorczyk [DH4,DH45] I showed that the widely applied linear model (BGe score, discussed in Section 6.1) is susceptible to incurring spurious feedback loops, which are a consequence of nonlinear regulation and autocorrelation in the data. We have propose a nonlinear generalization of this model, described below, and we have demonstrated that this approach successfully represses spurious feedback loops.

When the objective is to infer regulatory networks from time series, as is typically the case in systems biology, the restriction of the model to linear processes can result in the prediction of spurious feedback loops. Consider the simple example shown in Figure 6.15. The graph shows two

interacting nodes. Node X is a regulator of node Y, and it also has a regulatory feedback loop acting back on itself. Node Y is regulated by node X, but does not contain a feedback loop. The figure shows both the state space representation, i.e. the recurrent graph, and the corresponding dynamic Bayesian network. Note that the latter is a valid DAG obtained by the standard procedure of unfolding the state space graph in time, as illustrated in Figure 6.2. First assume that the data generation processes are linear and hence consistent with the BGe model assumption, e.g.:

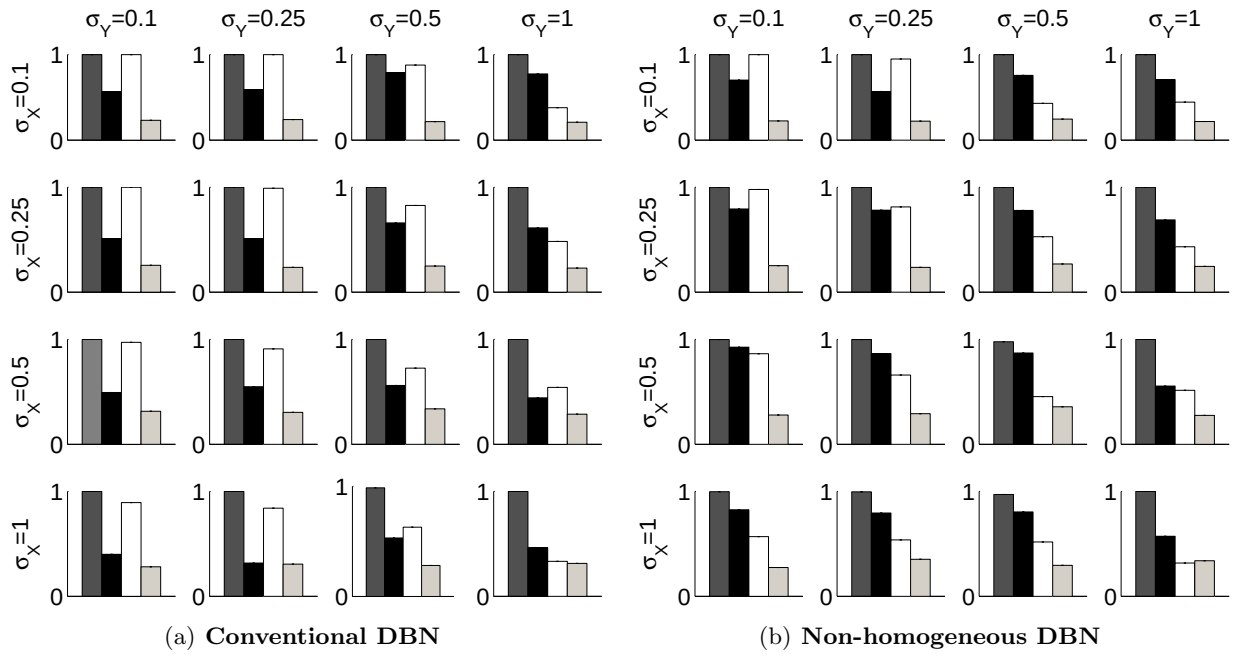
$$X(t+1) = X(t) + c + \sigma_x \cdot \phi_X(t); \quad Y(t+1) = w \cdot X(t) + m + \sigma_y \cdot \phi_Y(t) \tag{6.13}$$

where $w, m, c, \sigma_x, \sigma_y$ are constants, and $\phi_\cdot(.)$ are iid normally distributed random variables. Under fairly general regularity conditions, the marginal likelihood and, hence, the BGe score is a consistent estimator. This implies that the correct model structure will be learned as $T \to \infty$, where T is the data set size. Next, consider the scenario of a nonlinear regulatory influence that X exerts on Y:

$$X(t+1) = X(t) + c + \sigma_x \cdot \phi_X(t); \quad Y(t+1) = f(X(t)) + \sigma_y \cdot \phi_Y(t) \tag{6.14}$$

for some nonlinear function $f(.)$. This nonlinear function cannot be modelled with a linear Bayesian network based on the BGe model. Consequently, the prediction of $Y(t+1)$ from $X(t)$ will tend to be poor. Note that for sufficiently small noise levels, the $Y(t)$'s will exhibit a strong autocorrelation, by virtue of the autocorrelation of the $X(t)$'s, and the regulatory influence of $X(t)$ on $Y(t+1)$. If the latter regulatory influence cannot be learned owing to the linear restriction of our model, the next best explanation is a direct modelling of the autocorrelation between the $Y(t)$'s themselves. This autocorrelation corresponds to a feedback loop of Y acting back on itself in the state-space graph, or, equivalently, an edge from $Y(t)$ to $Y(t+1)$ in the dynamic Bayesian network; see Figure 6.15. The linear restriction of the Bayesian network model may therefore result in the prediction of spurious feedback loops and, hence, to the reconstruction of wrong network structures. Ruling out feedback loops altogether will not provide a sufficient remedy for this problem, as some nodes – X in the example above – will exhibit regulatory feedback loops (e.g. in molecular biology: transcription factors regulating their own transcription), and it is generally not known in advance where these nodes are.

In my work with Marco Grzegorczyk, we proposed and investigated a series of methods. The model in [DH12] is based on a mixture model, using latent variables to assign individual measurements to different classes. The practical inference follows the Bayesian paradigm and samples the network structure, the number of classes and the assignment of latent variables from the posterior distribution with MCMC, using the allocation sampler of Nobile and Fearnside [102] as an alternative to RJMCMC [54]. In [DH3,DH4,DH44,DH45], we increased the flexibility of the model by introducing different allocations for different nodes, so as to effectively approximate nonlinear regulatory processes by piecewise linear ones. However, we changed the assignment of data points to mixture components from a free allocation to a changepoint process. This effectively reduces the complexity of the latent variable space from exponential to polynomial complexity and incorporates our prior belief that, in a time series, adjacent time points are likely to be assigned to the same component.



(a) **Conventional DBN** (b) **Non-homogeneous DBN**

Figure 6.16: **Marginal edge posterior probabilities for the synthetic network.** Both panels are laid out as matrices, whose cells correspond to standard deviations σ_X and σ_Y of the noise in X and Y. All histograms show averages (means/standard deviations) from 20 independent data instantiations. The histograms show the posterior probabilities of the edges in the simple network of Figure 6.15, as obtained with a standard DBN using the BGe score (a) and the non-homogeneous DBN proposed in [DH4,DH44,DH45] (b). Each histogram contains 4 bars, which represent the average posterior probabilities of the 4 possible edges. Left: self-loop $X \to X$ (true); centre left: $X \to Y$ (true); centre right: self-loop $Y \to Y$ (false); right: $Y \to X$ (false). It is seen that the conventional linear homogeneous DBN (BGe score) has a high propensity for learning the spurious feedback loop $Y \to Y$, while the non-homogeneous DBN tends to learn an increased posterior probability of the correct edge $X \to Y$ (centre left bars).

In [DH6,DH7], we compared these two approaches systematically: the multiple changepoint process versus the free allocation mixture model. Conceptually, the former aims to relax the homogeneity assumption of DBNs, while the latter is more flexible and, in principle, more adequate for modelling nonlinear processes. However, in our study we discuss a potential theoretical disadvantage of the latter approach. In an empirical evaluation, we show that a model based on the multiple changepoint process achieves a systematically better performance than the free allocation model when inferring non-stationary gene regulatory processes from simulated and various real-world gene expression time series (macrophages challenged with viral infection, circadian regulation in *Arabidopsis thaliana*, and morphogenesis in *Drosophila melanogaster*). This justifies the approach taken in [DH3,DH4,DH44,DH45].

The approach in [DH3,DH4,DH44,DH45] is conceptually similar to the one discussed in Section 6.5. There are various differences, though: only the parameters rather than the structure may vary

between changepoints, there is a functional difference in the form of the priors on changepoints and parameters, and the underlying motivation/objective (improved nonlinear model flexibility) is different. To demonstrate the improvement in nonlinear modelling capability, I will provide a very simple illustration here, for demonstration purposes. Comprehensive comparative evaluation studies on more complex simulated and real-world data are available from the publications cited above.

Consider synthetic data generated from the simple network of Figure 6.15 according to the dynamics defined by the nonlinear state-space equations (6.14), with the nonlinear function $f(.) = sin(.)$. In [DH45] we generated 40 observations by applying equation (6.14) and setting the drift term $c = 2\pi/41$ to ensure that the complete period $[0, 2\pi]$ of the sinusoid is involved. Figure 6.16 shows the marginal posterior probabilities of the four possible edges in the two-node network of Figure 6.15 and the nonlinear state space process of (6.14). The results Figure 6.16(a) were obtained with the linear BGe model and show a clear propensity for inferring the spurious self-loop $Y \rightarrow Y$. Compare this with the results for the proposed non-homogeneous DBN, shown in Figure 6.16(b). Here, the spurious self-loop $Y \rightarrow Y$ is suppressed in favour of the correct edge $X \rightarrow Y$. There are two noise regimes in which the spurious self-loop $Y \rightarrow Y$ has a marginal posterior probability that is higher than or equal to that of the correct edge $X \rightarrow Y$. One noise regime is where both noise levels in X and Y are low (top left corner in panels (a) and (b) of Figure 6.16). Here, the autocorrelation of Y is so high that the spurious self-loop $Y \rightarrow Y$ is still favoured over the true edge $X \rightarrow Y$; this is a consequence of the fact that the functional dependence of $Y(t+1)$ on $X(t)$ is only learned approximately (namely approximated by a piecewise linear model induced by a multiple changepoint process). The second regime is where both noise levels are high (bottom right corners in panels (a) and (b) of Figure 6.16). High noise in Y blurs the functional dependence of $Y(t+1)$ on $X(t)$, while high noise in X leads to a high mis-classification of latent variables and, consequently, a deterioration of the model accuracy; this is a consequence of the fact that latent variables are not allocated individually, as in [DH12], but according to a changepoint process. However, in the majority of noise scenarios, the marginal posterior probability of the correct edge $X \rightarrow Y$ is significantly higher than that of the self-loop $Y \rightarrow Y$. This suggests that the proposed non-homogeneous DBN is successful at suppressing spurious feedback loops.

Our studies in [DH3,DH4,DH44,DH45] include a variety of further comparative evaluations on both synthetic and real-world data. We demonstrate the superiority of the proposed approach over the classical models BGe and BDe (see Section 6.1) as well as a competing nonlinear model [76]; we demonstrate the repression of potentially spurious feedback loops in regulatory networks related to macrophages challenged with viral infection; and we show that the gene regulatory network inferred from gene expression time series of nine circadian clock-regulated genes in *Arabidopsis thaliana* shows features that are consistent with the biological literature. We have also further improved the method by introducing a novel Bayesian clustering and information sharing scheme among nodes, which provides a mechanism for automatic model complexity tuning; see [DH3] for details.

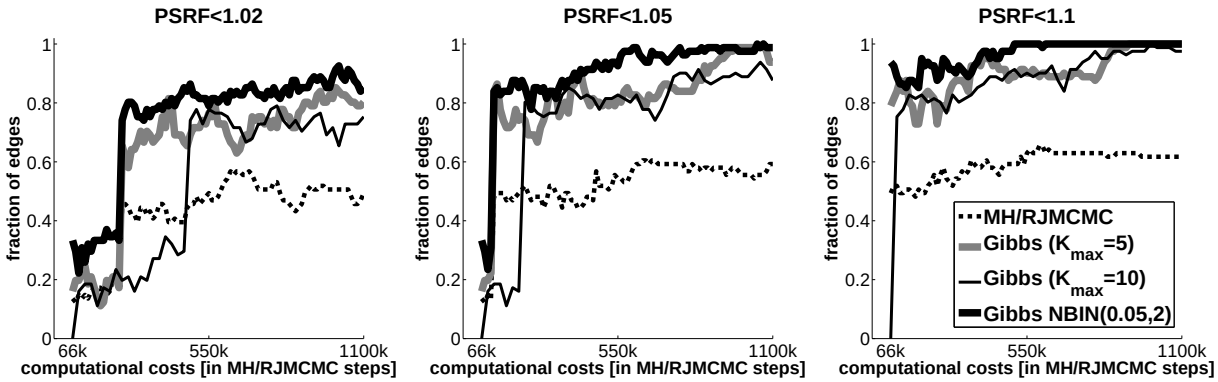


Figure 6.17: **Convergence diagnostics.** The graphs show the proportion of edges for which the potential scale redcution factor (PSRF) [49] lies below the indicated threshold, satisfying the respective convergence criterion. The horizontal axes represent simulation time, measured in terms of the equivalent number of MH/RJMCMC steps. Four MCMC schemes are compared. MH/RJMCMC: conventional Metropolis-Hastings based on single-edge proposal moves and RJMCMC for the changepoints. The other alternatives use the Gibbs sampling scheme proposed in [DH4] and described in Section 6.7, with three different priors for the changepoints: truncated Poisson priors on the number of components, truncated at the indicated value K_{max}, and the negative binomial distribution (NBIN) from (6.15).

6.7 Mixing and Convergence of MCMC

In [DH14] I demonstrated, in collaboration with Marco Grzegorczyk, how mixing and convergence of MCMC simulations for Bayesian learning of static Bayesian networks can be significantly improved. In [DH4], we discussed a significant improvement in MCMC convergence and mixing for non-homogeneous DBNs. In the present section, I use the following notation: $\mathcal{M}$ denotes the network structure, $\mathcal{K}_n$ denotes the number of components (number of changepoints plus 1) associated with node n, $\mathbf{V}_n$ is a vector of latent allocation variables assigning node n to its components, the vector $\mathbf{K} = (\mathcal{K}_1, \ldots, \mathcal{K}_N)$ and the matrix $\mathbf{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_N)$ combine the respective values for all nodes. To sample from the posterior distribution, $P(\mathcal{M}, \mathbf{V}, \mathbf{K}|\mathcal{D})$, all previous studies [83, 84, 118], including our own work in [DH44], follow the same procedure: to sample the network structure $\mathcal{M}$, they follow [51, 90] and apply structure MCMC with the Metropolis-Hastings algorithm [59], based on single-edge operations; to sample the latent variables $(\mathbf{V}, \mathbf{K})$, they follow [54] and apply reversible jump Markov chain Monte Carlo (RJMCMC), based on changepoint birth, death, and reallocation moves. In [DH4], we propose an improved scheme based on dynamic programming. The idea is to adapt the method proposed by Fearnhead [39] in the context of Bayesian mixture models to non-homogeneous DBNs of the form discussed in the previous two sections. Fearnhead [39] assumes that the changepoints occur at discrete time points, and he considers two priors for the changepoints. The first prior is based on a prior for the number of changepoints, and then a conditional prior on their positions. This corresponds exactly to $P(\mathcal{K}_n)$ and $P(\mathbf{V}_n|\mathcal{K}_n)$, as we used in the work summarized in the previous sections. The second prior is obtained from a point process on the positive and negative integers. The point process is specified by the probability mass

function $g(t)$ for the time between two successive points, for which a natural choice is the negative binomial distribution

$$g(t|a,p) = \binom{t-a}{a-1} p^a (1-p)^{t-a} \tag{6.15}$$

whose form is defined by two hyperparameters, a and p. The choice of this prior immediately imposes a prior distribution on the latent variables $\mathbf{V}_n$ without any conditioning on $\mathcal{K}_n$, $P(\mathbf{V}_n|\mathcal{K}_n) \rightarrow P(\mathbf{V}_n)$. For the remainder of this section, I use the generic notation $\tilde{\mathbf{V}} = (\tilde{\mathbf{V}}_1, \ldots, \tilde{\mathbf{V}}_N)$ to denote the latent variables induced by the changepoint prior. Depending on the form of the latter, we either have $\tilde{\mathbf{V}} = (\mathbf{V}, \mathbf{K})$ or $\tilde{\mathbf{V}} = \mathbf{V}$. Given a Bayesian mixture model for which the latent variables are of the form of one of the two changepoint processes discussed above, and the parameters can be integrated out in the likelihood, the changepoints can be sampled from the proper posterior distribution *exactly*, with a dynamic programming scheme. The computational complexity is quadratic in the number of observations T. To adapt this scheme to the inference of non-homogeneous DBNs, note that the Bayesian sampling of $P(\mathcal{M}, \tilde{\mathbf{V}}|\mathcal{D})$ can in principle follow a Gibbs sampling procedure, iteratively sampling the latent variables from $P(\tilde{\mathbf{V}}|\mathcal{M}, \mathcal{D})$, and a new network structure from $P(\mathcal{M}|\tilde{\mathbf{V}}, \mathcal{D})$. The first step can be accomplished with dynamic programming. However, given the comparatively high computational costs, the overall scheme is computationally inefficient if we follow [83, 84, 118] and [DH44] and stick to a structure MCMC step for updating $\mathcal{M}$, i.e. if we follow a computationally expensive complete Gibbs step for sampling from $P(\tilde{\mathbf{V}}|\mathcal{M}, \mathcal{D})$ by a computationally cheap Metropolis-Hastings-within-Gibbs step for incomplete sampling from $P(\mathcal{M}|\tilde{\mathbf{V}}, \mathcal{D})$. To resolve this issue, we adapted the sampling scheme proposed in [45], Eq. (10). Note that the network structure $\mathcal{M}$ is defined by the complete set of parent sets $\{\pi_n\}_{1 \leq n \leq N}$. Having sampled $\tilde{\mathbf{V}} = (\tilde{\mathbf{V}}_1, \ldots, \tilde{\mathbf{V}}_N)$ from $P(\tilde{\mathbf{V}}|\mathcal{M}, \mathcal{D})$ in the previous Gibbs step, we now sample $\mathcal{M}$ from $P(\mathcal{M}|\tilde{\mathbf{V}}, \mathcal{D})$ by sampling, for all nodes $n \in \{1, \ldots, N\}$ in turn, new parent configurations $\{\pi_n\}$ from

$$P(\pi_n|\mathcal{D}, \tilde{\mathbf{V}}_n) \;=\; \Psi^\dagger(\mathcal{D}_n^{\pi_n}[\tilde{\mathbf{V}}_n]) / \sum\nolimits_{\tilde{\pi}_n} \Psi^\dagger(\mathcal{D}_n^{\tilde{\pi}_n}[\tilde{\mathbf{V}}_n]) \tag{6.16}$$

where $\Psi^\dagger(\mathcal{D}_n^{\pi_n}[\tilde{\mathbf{V}}_n])$ is a local score corresponding to the marginal likelihood of a configuration defined by node n and parents π_n. Equation (6.16) entails a complete enumeration over all parent configurations, which is computationally expensive. In [DH4] we therefore investigated under which conditions the dynamic programming scheme for exact sampling of $\tilde{\mathbf{V}}$ from $P(\tilde{\mathbf{V}}|\mathcal{M}, \mathcal{D})$ achieves an overall gain in computational efficiency. We also investigated how critically the computational costs depend on the prior distribution for the changepoints. An illustration, obtained from gene expression time series of nine circadian genes in *Arabidopsis thaliana*, is shown in Figure 6.17. The results from a comprehensive comparative evaluation study, as well as a detailed mathematical derivation of the algorithm, are available from [DH4].

Chapter 7

Modelling Transcriptional Regulation

Understanding the mechanisms of gene transcriptional regulation through analysis of high-throughput postgenomic data is one of the central problems of computational systems biology. Various approaches have been proposed, but most of them fail to address at least one of the following objectives: (1) allow for the fact that transcription factors are potentially subject to post-transcriptional regulation; (2) allow for the fact that transcription factors co-operate as a functional complex in regulating gene expression, and (3) provide a model and a learning algorithm with manageable computational complexity. The objective of my work with Kuang Lin [DH8] was to develop and test a method that addresses these three issues. Our approach aims to integrate gene expression profiles with transcription factor binding data, such as binding motifs in promoter regions or p-values from immunoprecipitation experiments. Our model is a mixture of factor analyzers, in which the latent variables correspond to different transcription factors, grouped into complexes or modules. Inference is carried out in a Bayesian framework, using the Variational Bayesian Expectation Maximization (VBEM) algorithm for approximate inference of the posterior distributions of the model parameters, and estimation of a lower bound on the marginal likelihood for model selection. We carried out a comparative evaluation with other state of the art methods to assess the accuracy of transcription factor activity profile reconstruction and regulatory network inference.

7.1 Overview

Quantitative modelling of the regulatory networks of the cell is one of the central challenges in computational biology. One of the most important cellular regulation mechanisms is at the transcriptional level, where the expression of a gene is controlled by the binding of diverse regulatory proteins called transcription factors (TFs) to specific DNA sequences in the promoter region of the gene. Transcriptional gene regulation is a complex process that utilizes a network of interactions. These networks control the expression levels of thousands of genes as part of diverse biological processes such as the cell cycle, embryogenesis, host-pathogen interactions and circadian rhythms. Determining accurate models for TF-gene regulatory interactions is thus an important challenge of computational systems biology. Most recent studies of transcriptional regulation can be placed broadly in one of three categories.

Approaches in the first class attempt to build quantitative models to associate gene expression levels, as typically obtained from microarray experiments, with putative binding motifs on the gene promoter sequences. Bussemaker et al. [20] and Conlon et al. [29] propose a linear regression model for the dependence of the log gene expression ratio on the presence of regulatory sequence motifs. Beer and Tavazoie [11] cluster gene expression profiles in a preliminary data analysis based on correlation, and then apply a Bayesian network classifier to predict cluster membership from sequence motifs. Phuong et al. [106] use multivariate decision trees to find motif combinations that define homogeneous groups of genes with similar expression profiles. Segal et al. [131] cluster genes with a probabilistic generative model that systematically integrates gene expression profiles with regulatory sequence motifs.

A shortcoming of the methods in the first class is that the activities of the TFs are not included in the model. This limitation is addressed by models in the second class, which predict gene expression levels from both binding motifs in promoter sequences and the expression levels of putative regulators. Middendorf et al. [99, 98] approach this problem as a binary classification task to predict up- and down-regulation of a gene from a combination of a motif presence/absence indication and the discrete transcriptional state of a putative regulator. The bi-dimensional regression trees of Ruan and Zhang [120] are based on a similar idea, but avoid the information loss inherent in the binary gene expression discretization. My work with Adriano Werhli [DH16], summarized in the previous chapter, falls into the same class in that regulatory networks of genes are inferred from transcriptional profiles of genes and their putative regulators complemented with prior knowledge about binding motifs in promoter sequences.

Transcriptional regulation is influenced by TF *activities*, that is the concentration of the TF subpopulation capable of DNA binding. The methods in the second class approximate the activities of TFs by their gene expression levels. However, TFs are frequently subject to post-translational modifications, which may affect their DNA binding capability. Consequently, gene expression levels

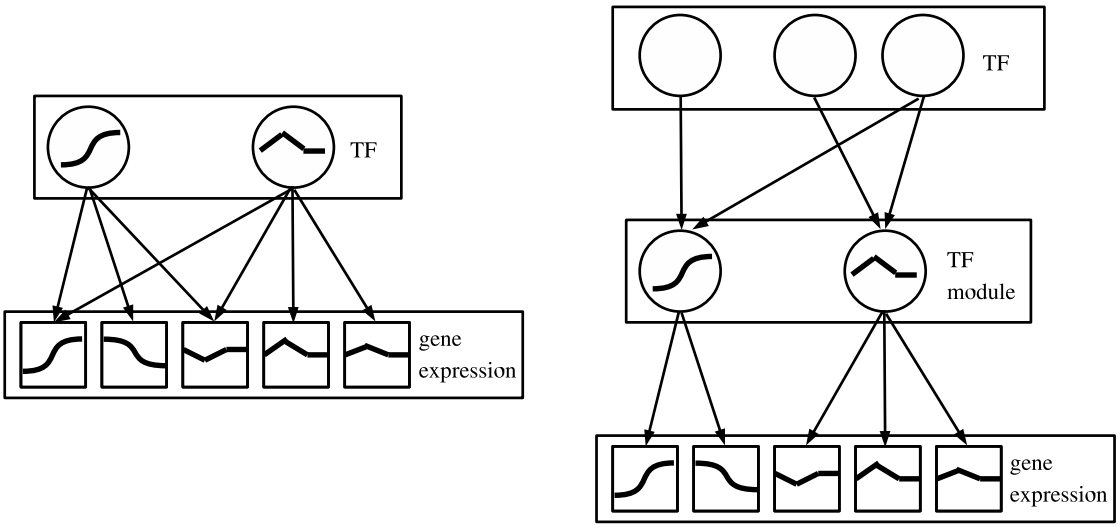


Figure 7.1: **Transcriptional regulatory network.** The left panel shows a transcriptional regulatory network in the form of a bipartite graph, in which a small number of transcription factors (TFs), represented by circles, regulate a large number of genes (represented by squares) by binding to their promoter regions. The black lines in the square boxes indicate gene expression profiles, that is, gene expression values measured under different experimental conditions or for different time points. The black lines in the circles represent TF activity profiles, i.e. the concentrations of the TF subpopulation capable of DNA binding. Note that these TF activity profiles may be unobserved owing to post-translational modifications, and should hence be included as hidden or latent variables in the statistical model. The right panel shows a more accurate representation of transcriptional regulation that allows for the cooperation of several TFs forming functional complexes; this complex formation is particularly common in higher eukaryotes.

of TFs may contain only limited information about their actual activities. The methods in the third class address this shortcoming by treating TFs as latent or hidden components. The regulatory system is modelled as a bipartite network, as shown in the left panel of Figure 7.1, in which high-dimensional output data are driven by low-dimensional regulatory signals. The high-dimensional output data correspond to the expression levels of a large number of regulated genes. The regulators correspond to a comparatively small number of TFs, whose activities are unknown. Various authors have applied latent variable models like principal component analysis (PCA), factor analysis (FA), and independent component analysis (ICA) to determine a low-dimensional representation of high-dimensional gene expression profiles [87, 110]. However, these approaches provide only a phenomenological modelling of the observed data, and the hidden components do not correspond to identified TFs. Liao et al. [86] and Kao et al. [74] address this shortcoming by including partial prior knowledge about TF-gene interactions, as obtained from Chromatin Immunoprecipitation (ChIP) experiments [56] or TF binding motif finding algorithms [5, 67]. Their network component analysis (NCA) is equivalent to a constrained maximum likelihood procedure in the presence of Gaussian noise and independent hidden components; the latter represent the TF activities. A ma-

jor limitation of NCA is the fact that the constraints on the connectivity pattern of the bipartite network are rigid, which does not allow for the noise intrinsic to immunoprecipitation experiments or sequence motif detection. Sabatti and James [122] and Sanguinetti et al. [125] address this shortcoming by proposing an approach based on Bayesian factor analysis, in which prior knowledge about TF-gene interactions naturally enters the model in the form of a prior distribution on the elements of the loading matrix. Pournara and Wernisch [108] propose an alternative approach based on maximum likelihood, where the loading matrix is orthogonally rotated towards a target matrix of *a priori* known TF-gene interactions. All three approaches simultaneously reconstruct the structure of the bipartite regulatory network – represented by the loading matrix – and the TF activity profiles – represented by the hidden factors – from gene expression data and (noisy) prior knowledge about TF-gene interactions. In a generalization of these approaches, Shi et al. [132] have introduced a further latent variable to indicate whether a TF is transcriptionally or post-transcriptionally regulated.

Contrary to the methods in the first two classes, the methods in the third class do not incorporate interaction effects between TFs, though. This is a major limitation, since especially in higher eukaryotes transcription factors co-operate as a functional complex in regulating gene expression [112, 154]. Boulesteix and Strimmer [16] allow for this complex formation by proposing a latent variable model in which the latent components correspond to groups of TFs. However, their partial-least squares (PLS) approach does not provide a probabilistic model and hence, like NCA, does not allow for the noise inherent in TF binding profiles from immunoprecipitation experiments or sequence motif detection schemes.

In my work with Kuang Lin [DH8] we aimed to combine the advantages of the methods in the three classes summarized above. Like the approaches in the third class, our method is a latent variable model that allows for the fact that owing to post-translational modifications the true TF activities are unknown. Similar to the approaches of the first two classes, our model explicitly incorporates interactions among TFs. Inspired by [16], we aim to group individual TFs into TF modules, as illustrated in the right panel of Figure 7.1. To allow for the noise inherent in both gene expression levels and TF binding profiles, we use a proper probabilistic generative model, like [122, 125]. Our work is based on the work of Beal [9]. We apply a mixture of factor analyzers model, in which each component of the mixture corresponds to a TF complex composed of several TFs. This approach allows for the fact that TFs are not independent. By explicitly including this in our model we gain a more parsimonious representation with fewer parameters, and hence more stable inference. To further improve the robustness of this approach, we pursue inference in a Bayesian framework, which includes a model selection scheme for estimating the number of TF complexes.

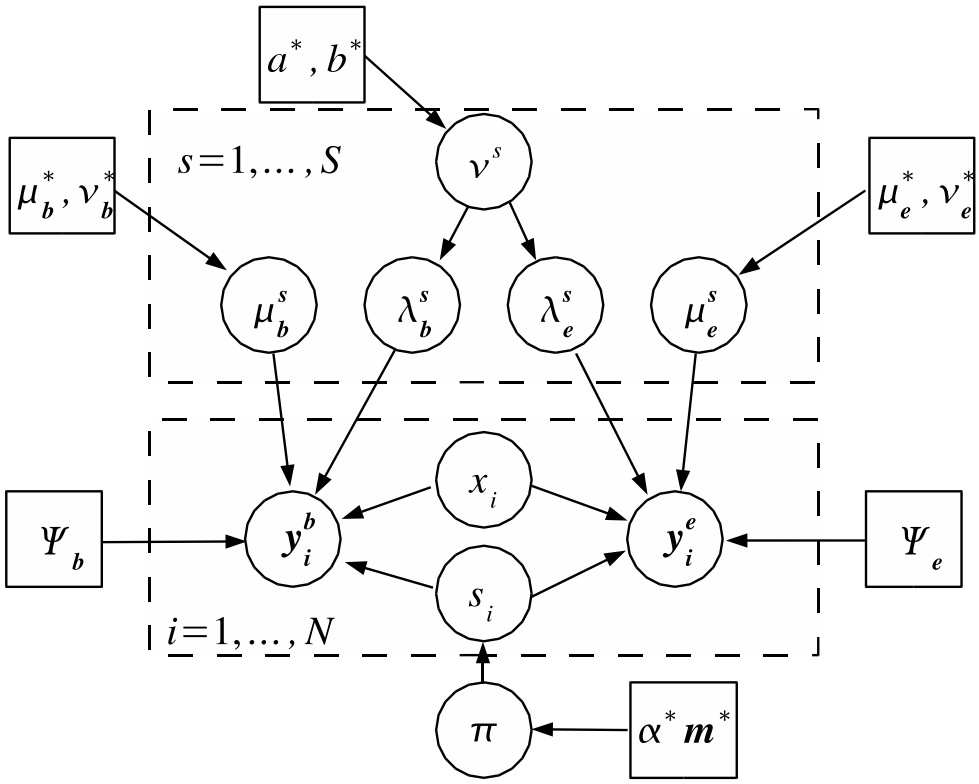


Figure 7.2: **Bayesian mixture of factor analyzers (MFA) model applied to transcriptional regulation.** The figure shows a probabilistic independence graph of the Bayesian mixture of factor analyzers (MFA) model proposed in [DH8]. Variables are represented by circles, and hyperparameters are shown as square boxes in the graph. S components (factor analyzers), each with their own parameters $\boldsymbol{\lambda}^s = [\boldsymbol{\lambda}_e^s, \boldsymbol{\lambda}_b^s]$ and $\boldsymbol{\mu}^s = [\boldsymbol{\mu}_e^s, \boldsymbol{\mu}_b^s]$, are used to model the expression profiles $\mathbf{y}_i^e$ and TF binding profiles $\mathbf{y}_i^b$ of $i = 1, \ldots, N$ genes. The factor loadings $\boldsymbol{\lambda}^s$ have a zero-mean Gaussian prior distribution, whose precision hyperparameters ν^s are given a gamma distribution determined by a^* and b^*. The analyzer displacements $\boldsymbol{\mu}_e^s$ and $\boldsymbol{\mu}_b^s$ have Gaussian priors determined by the hyperparameters $\{\boldsymbol{\mu}_e^*, \boldsymbol{\nu}_e^*\}$ and $\{\boldsymbol{\mu}_b^*, \boldsymbol{\nu}_b^*\}$, respectively. The indicator variables $s_i \in \{1, \ldots, S\}$ select one out of S factor analyzers, and the associated latent variables or factors x_i have Normal prior distributions. The indicator variables s_i are given a multinomial distribution, whose parameter vector $\boldsymbol{\pi}$, the so-called mixture proportions, have a conjugate Dirichlet prior with hyperparameters $\alpha^* \mathbf{m}^*$. $\boldsymbol{\Psi}_e$ and $\boldsymbol{\Psi}_b$ are the diagonal covariance matrices of the Gaussian noise in the expression and binding profiles, respectively. A dashed rectangle denotes a plate, that is an iid repetition over the genes $i = 1, \ldots, N$ or the mixture components $s = 1, \ldots, S$, respectively. The biological interpretation of the model is as follows. $\boldsymbol{\mu}_b^s$ represents the composition of the sth transcriptional module, that is, it indicates which TFs bind cooperatively to the promoters of the regulated genes. $\boldsymbol{\lambda}_b^s$ allows for perturbations that result *e.g.* from the temporary inaccessibility of certain binding sites or a variability of the binding affinities caused by external influences. $\boldsymbol{\mu}_e^s$ is the background gene expression profile. $\boldsymbol{\lambda}_e^s$ represents the activity profile of the sth transcriptional module, which modulates the expression levels of the regulated genes. x_i describes the gene-specific susceptibility to transcriptional regulation, that is, to what extent the expression of the ith gene is influenced by the binding of a transcriptional module to its promoter. A complete description of the model can be found in [DH8].

7.2 Method

The model we proposed in [DH8] is shown in Figure 7.2. A complete description can be found in our paper. In this high-level overview I will only outline the central ideas of variational Bayesian inference, which we applied. Denote by $\mathbf{X} = (X^1, \ldots, X^T)$ the latent variables, which are related to the unknown TF activity profiles and TF complex memberships, and by $\boldsymbol{\theta}$ the model parameters. The latter are treated as random variables for which some prior distribution $P(\boldsymbol{\theta})$ is defined. The objective of Bayesian inference is to infer the posterior distribution $p(\boldsymbol{\theta}, \mathbf{X}|\mathbf{Y}, \mathcal{M})$ from the data $\mathbf{Y} = \{Y^1, \ldots, Y^T\}$ for some model $\mathcal{M}$, and to decide on the best model $\mathcal{M}$ on the basis of the marginal likelihood $P(\mathbf{Y}|\mathcal{M})$. Recall that the data $\mathbf{Y}$ includes transcriptional profiles (from microarrays) and TF bindings indications (from immunoprecipitation assays). In the context of our work [DH8], the model selection task is to decide on the number of components in the mixture. Unfortunately, neither $P(\boldsymbol{\theta}, \mathbf{X}|\mathbf{Y}, \mathcal{M})$ nor $P(\mathbf{Y}|\mathcal{M})$ can be computed in closed form. The objective of variational inference is to approximate both on the basis of an analytically tractable model distribution $Q(\mathbf{X}, \boldsymbol{\theta})$. Define

$$\mathcal{F} \;=\; \int Q(\mathbf{X}, \boldsymbol{\theta}) \log \left(\frac{P(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Y}|\mathcal{M})}{Q(\mathbf{X}, \boldsymbol{\theta})} \right) \mathrm{d}\mathbf{X}\mathrm{d}\boldsymbol{\theta} \tag{7.1}$$

It is easy to show that $\mathcal{F}$ can be decomposed into the following form:

$$\mathcal{F} \;=\; \log P(\mathbf{Y}|\mathcal{M}) - \mathrm{KL}\{Q(\mathbf{X}, \boldsymbol{\theta}) || P(\boldsymbol{\theta}, \mathbf{X}|\mathbf{Y}, \mathcal{M})\} \tag{7.2}$$

which is the difference of the log marginal likelihood and the Kullback-Leibler divergence between the model distribution $Q(\mathbf{X}, \boldsymbol{\theta})$ and the unknown true posterior distribution $P(\boldsymbol{\theta}, \mathbf{X}|\mathbf{Y}, \mathcal{M})$:

$$\mathrm{KL}\{Q(\mathbf{X}, \boldsymbol{\theta}) || P(\boldsymbol{\theta}, \mathbf{X}|\mathbf{Y}, \mathcal{M})\} \;=\; \int Q(\mathbf{X}, \boldsymbol{\theta}) \log \left(\frac{Q(\mathbf{X}, \boldsymbol{\theta})}{P(\boldsymbol{\theta}, \mathbf{X}|\mathbf{Y}, \mathcal{M})} \right) \mathrm{d}\mathbf{X}\mathrm{d}\boldsymbol{\theta} \tag{7.3}$$

From information theory it is known that the Kullback-Leibler divergence, which is a measure of the difference between two distributions, is non-negative; see for instance Chapter 2 in [DH1]. This implies that $\mathcal{F}$ is a lower bound on the marginal likelihood $\log P(\mathbf{Y}|\mathcal{M})$, with a difference given by the the Kullback-Leibler divergence $\mathrm{KL}\{Q(\mathbf{X}, \boldsymbol{\theta}) || P(\boldsymbol{\theta}, \mathbf{X}|\mathbf{Y}, \mathcal{M})\}$. The objective of variational Bayesian inference is to numerically maximize $\mathcal{F}$. This gives the best approximation to the true posterior distribution from the functional family Q, while simultaneously $\mathcal{F}$ gives the best possible approximation to the marginal likelihood.

To apply the concept of variational learning to the model we proposed in [DH8], the model distribution is assumed to factorize into separate contributions from the parameters and latent variables: $Q(\boldsymbol{\theta}, \mathbf{X}) = Q(\boldsymbol{\theta})Q(\mathbf{X})$. The variational learning algorithm then iteratively maximizes the functional $\mathcal{F}$ with respect to the free distributions $Q(\boldsymbol{\theta})$ and $Q(\mathbf{X})$. Given a fixed distribution of the parameters $Q^{(t)}(\boldsymbol{\theta})$, $\mathcal{F}$ is maximized with respect to $Q(\mathbf{X})$ by setting to zero the following functional derivative:

$$\frac{\delta}{\delta Q(\mathbf{X})} \left(\mathcal{F} + \xi \left[\int Q(\mathbf{X}) \mathrm{d}\mathbf{X} - 1 \right] \right) \;=\; 0 \tag{7.4}$$

Table 7.1: **Overview of methods.** The table shows an overview of the methods with which we compared the model we proposed in [DH8].

PLS	The partial least squares approach proposed in [16]. Note that the method treats TF-gene interactions as fixed constants that cannot be changed in light of the gene expression data. Hence, this approach cannot be used for network reconstruction and was only applied for reconstructing the TF activity profiles.
FA	Maximum likelihood factor analysis, effected with the EM algorithm [50] and a subsequent varimax rotation [70] of the loading matrix towards maximum sparsity as proposed in [108].
BFA-Gibbs	Bayesian factor analysis [122], trained with Gibbs sampling. The TF regulatory network is obtained from the posterior expected loading matrix.
MFA-VBEM	Our mixture of factor analyzers model, shown in Figure 7.2 and discussed in [DH8], trained with variational Bayesian Expectation Maximization. The approach is based on the work of Beal [9].

where ξ is a Lagrange multiplier resulting from the normalization constraint. Equation (7.4) has the formal closed-form solution:

$$Q^{(t+1)}(\mathbf{X}) \;=\; \frac{1}{\mathcal{Z}_{\mathbf{X}}} \exp\left[\int \mathrm{d}\boldsymbol{\theta} Q^{(t)}(\boldsymbol{\theta}) \log P(\mathbf{Y}, \mathbf{X}|\boldsymbol{\theta}, \mathcal{M})\right] \tag{7.5}$$

where $\mathcal{Z}_{\mathbf{X}}$ denotes a normalization constant. Likewise, given a fixed distribution of the latent variables $Q^{(t+1)}(\mathbf{X})$, $\mathcal{F}$ is maximized with respect to $\mathrm{Q}(\boldsymbol{\theta})$ by setting to zero the following functional derivative:

$$\frac{\delta}{\delta Q(\boldsymbol{\theta})}\left(\mathcal{F} + \xi\left[\int Q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} - 1\right]\right) \;=\; 0 \tag{7.6}$$

which has the formal closed form solution

$$Q^{(t+1)}(\boldsymbol{\theta}) \;=\; \frac{1}{\mathcal{Z}_{\boldsymbol{\theta}}} P(\boldsymbol{\theta}|\mathcal{M}) \exp\left[\int \mathrm{d}\mathbf{X} Q^{(t+1)}(\mathbf{X}) \log P(\mathbf{Y}, \mathbf{X}|\boldsymbol{\theta}, \mathcal{M})\right] \tag{7.7}$$

Again, $\mathcal{Z}_{\boldsymbol{\theta}}$ is a normalization constant. For a derivation of these results, see [9]. For distributions of the exponential family, which includes our model in [DH8], the integrals in equations (7.5) and (7.7) have a closed-form solution, as shown in [9]. In analogy to the Expectation Maximization (EM) algorithm, the variational learning algorithm follows an iterative adaptation procedure including the following two steps. Variational E-step: Given the distribution of the parameters $Q^{(t)}(\boldsymbol{\theta})$, where t indicates the iteration number, obtain a new distribution of the latent variables $Q^{(t+1)}(\mathbf{X})$ by application of equation (7.5). Variational M-step: Given the distribution of the latent variables $Q^{(t+1)}(\mathbf{X})$, obtain a new distribution of the parameters $Q^{(t+1)}(\boldsymbol{\theta})$ by application of equation (7.7). This procedure, called the Variational Bayesian Expectation Maximization (VBEM) algorithm, is repeated until a stationary point of $\mathcal{F}$ is reached.

Table 7.2: **Reconstruction of TF complex activity profiles.** The table shows the mean absolute correlation coefficient between the true and inferred activity profiles, averaged over 6 simulated activity profiles. N1, N2 and N3 refer to the three noise levels of $\mathbf{e}_i \sim \mathcal{N}(0, 0.25\mathbf{I})$, $\mathcal{N}(0, 0.5\mathbf{I})$ and $\mathcal{N}(0, \mathbf{I})$. L1, L2, and L3 refer to the expression profile lengths being 10, 20 and 40. Three methods have been compared: the partial least squares (PLS) approach proposed in [16]; Bayesian factor analysis (BFA) with Gibbs sampling [122]; and the MFA model trained with VBEM, proposed in [DH8].

Method	Lengths	N1	N2	N3
PLS		0.53	0.52	0.52
BFA	L1	0.92	0.89	0.78
MFA		0.88	0.83	0.71
PLS		0.52	0.51	0.52
BFA	L2	0.83	0.72	0.72
MFA		0.95	0.85	0.71
PLS		0.52	0.51	0.52
BFA	L3	0.90	0.73	0.67
MFA		0.98	0.94	0.63

Note that VBEM and MCMC are alternative paradigms for approximate Bayesian inference. While posterior estimates obtained with MCMC are, under certain regularity conditions (ergodicity), consistent, VBEM is known to systematically underestimate the posterior uncertainty; see e.g. [14] and Figure 5.2 in [82]. On the other hand, the computational costs for MCMC are usually substantially higher than for VBEM, and monitoring convergence is more straightforward for VBEM than for MCMC.

7.3 Findings

Reconstruction of TF activity profiles

Since TF activity profiles are not available for real data, we used synthetic data from a simulation study to evaluate the profile reconstruction performance of the model. Details of these simulations can be found in [DH8]. We compared the proposed MFA-VBEM model with the partial least-squares (PLS) approach of Boulesteix and Strimmer [16], and with the Bayesian factor analysis model using Gibbs sampling (BFA-Gibbs), as proposed in [122]. An overview of the models is shown in Table 7.1.

Table 7.2 shows a comparison of the reconstruction accuracy in terms of the mean absolute Pearson correlation between the true and estimated TF module activity profiles. It is seen that BFA-Gibbs and the proposed MFA-VBEM scheme consistently outperform PLS. The comparatively

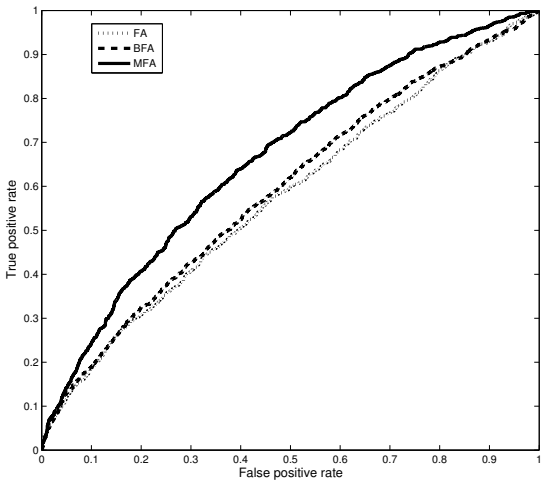


Figure 7.3: **Out-of-sample TF regulatory network reconstruction for yeast.** The figure shows ROC (Receiver Operating Characteristic) curves obtained for *S. cerevisiae* with three different methods: (1) solid line: the MFA-VBEM method we proposed in [DH8], adapted from [9]; (2) dashed line: the Bayesian FA model with Gibbs sampling, as proposed in [122]; and (3) dotted line: maximum likelihood FA trained with the EM algorithm [50] and a subsequent varimax rotation [70] of the loading matrix towards maximum sparsity, as proposed in [108].

poor performance of PLS, which has been independently reported in [108], is a consequence of the fact that PLS lacks any mechanism to deal with the noise inherent in the TF binding profiles. This shortcoming is addressed by both BFA-Gibbs and MFA-VBEM.

A comparison between BFA-Gibbs and MFA-VBEM shows that BFA-Gibbs tends to outperform MFA-VBEM when the expression profiles are short (length L1) or when the noise level is high (N3). This could be a consequence of the different inference schemes ("VBEM" versus "Gibbs"). Short expression profiles and high noise levels lead to diffuse posterior distributions of the parameters. Variational learning – as opposed to Gibbs sampling – is known to lead to a systematic underestimation of the posterior variation [14], which could be a disadvantage here. However, MFA-VBEM consistently outperforms BFA-Gibbs on the longer expression profiles with length L2 and L3, and the lower noise levels N1 and N2. It appears that this improvement in the performance is a consequence of the more parsimonious model that results when allowing for the fact that TFs are not independent, which leads to greater robustness of inference and reduced susceptibility to overfitting.

Regulatory network reconstruction

For evaluating the inference of transcriptional regulation in real organisms, we chose gene expression and TF binding data from the widely used model organism *Saccharomyces cerevisiae* (baker's yeast). We used the gene expression data from [101] and the TF binding profiles from YeastTract [141]. YeastTract[1] provides a comprehensive database of transcriptional regulatory associations in *S. cerevisiae*. Our combined data set included the expression levels of 5464 genes under 214 experimental conditions and binary TF binding patterns associating these genes with 169 TFs.

For the network prediction task, we trained the models on only 80% of the *S. cerevisiae* genes, and used an independent test set containing a randomly selected subset of 20% of the genes to estimate the out-of-sample network prediction accuracy. Note that for the genes in the test set, only the expression profiles were made available, while the corresponding TF binding connections were held back. The task was to predict these TF binding connections from the gene expression data, using the (average) TF activity profiles inferred from the training set. For a more comprehensive description of the evaluation, see [DH8]. The results are shown in Figure 7.3. None of the ROC curves is particularly good, which indicates that there is substantial room for further methodological innovation. However, the method we proposed in [DH8] clearly outperforms the competing approaches. This improved performance is a consequence of the more parsimonious model that results from modelling the fact that TFs are non-independent, but act via complexes; see Figure 7.1. On the contrary, there is nothing in the competing approaches that would inform the model *a priori* that once a group of TFs are found to form a module, their interaction patterns with the regulated genes should be the same. Instead, these interaction strengths have to be learned separately for each TF. This leads to a less parsimonious and partially redundant model, which is less robust and more susceptible to overfitting.

[1]Publicly available from `http://www.yeastract.com`.

Chapter 8

Outlook:
From Molecular Biology to Ecology

The complexity of ecosystems is staggering, with hundreds or thousands of species interacting in a number of ways from competition and predation to facilitation and mutualism. Understanding the networks that form the systems is of growing importance, e.g. to understand how species will respond to climate change, or to predict potential knock-on effects of a biological control agent. In the present chapter, I summarize my work in [DH5] and address the question to what extent the methods discussed in the previous chapters, used to reconstruct regulatory networks in molecular biology, can be adapted to infer species interaction networks in ecology, and which methodological hurdles have to be overcome.

8.1 Introduction

Darwin's description of a tangled bank describes the everyday complexity of ecology that we overlook at our peril. Tampering with the population of one species can cause surprising and dramatic changes in the populations of others [28, 64]. Altering pressures to which ecosystems are exposed can drive them to alternative states [12] or catastrophic failure [133]. Understanding and predicting how ecosystems will respond to change requires untangling the tangled bank and is of enormous importance during a period of rapid global change. Yet such a task can seem impossible given the enormous complexity of ecological systems and the excruciating fieldwork needed to quantify even the simplest of foodwebs [69, 97].

Molecular biology		Ecology
Genes	↔	Species
Expression levels	↔	Population densities
Gene regulation	↔	Species interactions
Different conditions	↔	Different environments

Table 8.1: **Comparison between molecular biology and ecology.** The table shows the formal correspondences between molecular systems biology and ecological.

The current approaches typically taken to learn the structure of a species interaction network are based on minute observations and detailed field work. For instance, the information theoretic summary statistics proposed in [15] were applied to the plant-pollinator interaction networks obtained in [96, 148]. These studies entailed detailed observations of how often a particular plant was visited by a particular pollinator, for all pollinators and plants in turn. This process is laborious and error-prone. More importantly, it is restricted to specific kinds of interactions. The interactions between pollinators and their host plants are amenable to direct observation. However, other types of species interactions, like competition for resources, facilitation[1] and mutualism[2], are not, and might not even be clearly defined from the outset.

However, information about ecological interactions should be evident in a range of ecological data that are currently available. For example, time-series of the populations of multiple species present in a study site should allow identification of important interactions, and similarly the spatial patterns of coincidence of species should contain information about the interactions among these species, potentially at a range of scales. In my joint work with Ali Faisal, Frank Dondelinger and Colin Beale [DH5], we therefore investigated to what extent network reconstruction methods currently applied in molecular systems biology, like those described in Chapter 6, can be applied to reconstruct interaction networks *in silico* from species abundance counts. A comparison of the formal correspondences can be found in Table 8.1.

8.2 Method

In the most general case, our aim in describing an ecological network is to model all the interactions between and among species and their environment. It is convenient to think of this network as a 'graph' (e.g. Fig. 8.4), describing species as the 'nodes' within the graph, and interactions as the links or 'edges' that join the nodes. To identify and infer these graphs we selected four widely used methods for network recovery in postgenomic data analysis: Graphical Gaussian models [126, 127], L1-regularized regression (LASSO) [142, 147], sparse Bayesian regression [119, 143] and Bayesian

[1] Ecological facilitation describes how an organism profits from the presence of another.

[2] Mutualism is the way two organisms biologically interact where each individual derives a fitness benefit.

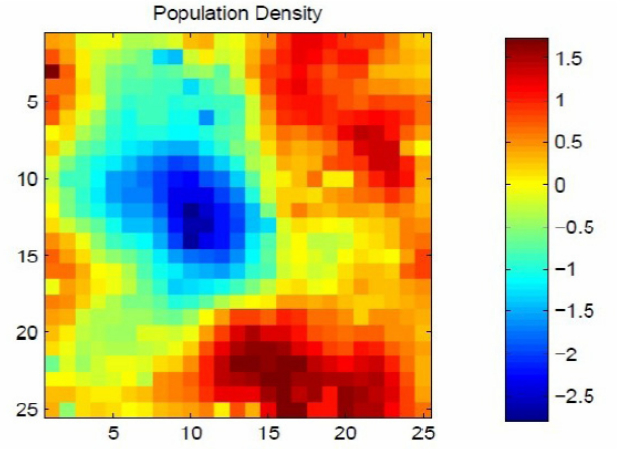


Figure 8.1: **Illustration of spatial auocorrelation.** The colour indicates the population density of a selected species, depending on the grid location. This is the result of a simulation study that combined a niche model [153] with a stochastic population model [79, Chapter 8] in a 2-dimensional 25-by-25 lattice and included predator-prey interactions in the Lotka-Volterra form. The effect of spatial autocorrelation is evident: densities at nearby locations are more similar than densities at more distant locations.

networks (as discussed in Chapter 6). For a review of these methods, see [DH5].

A particular methodological complication, which we did not have to address for the molecular data, is spatial autocorrelation. This phenomenon, that observations at nearby locations are more similar than observations at more distant locations, as illustrated in Figure 8.1, is nearly ubiquitous in ecology and can have a strong impact on statistical inference [30, 85]. In our case, spatial autocorrelation could lead to the identification of spurious interactions as a mere consequence of two species co-occurring in similar geographical regions. Where possible, we applied an autoregressive approach similar to that of [4] to incorporate potential spatial autocorrelation into the models. To this end, we computed the average population at neighbouring cells, weighted inversely proportional to the distance of the neighbours. We call this the autocorrelation variable:

$$a = \frac{\sum_{i=1}^{K} \omega_i x_i}{\sum_{i=1}^{K} \omega_i} \tag{8.1}$$

where K is the number of neighbours that we are considering (usually $K = 4$), x_i is the population density at neighbour i, and ω_i is the weight given to that neighbour, which is inversely proportional to the Euclidean distance of the neighbour. A slight subtlety when working with real world data that is not distributed in a regular grid is to work out which neighbouring locations to consider. In this work, we have opted for the closest neighbours by Euclidean distance. For Bayesian networks, we connect each node to a parent node whose value is given by (8.1), i.e. a representation of the spatial neighbourhood.[3] An illustration can be found in Figure 8.2(b). In this way the observation status

[3]The incoming edge from the parent node is enforced and excluded from the fan-in count.

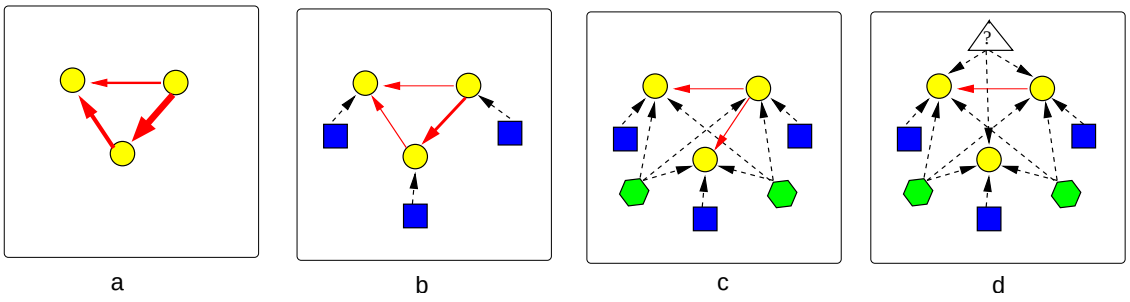


Figure 8.2: **Illustration of the improved method for ecological network reconstruction.** *Panel (a)* illustrates a species interaction network modelled by a Bayesian network. Yellow circles represent species (nodes), red arrows present species interactions (edges). Networks inferred from species abundance or population density data alone tend to contain many spurious interactions. In [DH5], we investigated three methods for suppressing spurious interactions, illustrated in the panels on the right. *Panel (b): Allowing for spatial autocorrelation.* Each node is hard-wired to an indicator node (blue square) that represents, via equation (8.1), the average population density in the spatial neighbourhood. *Panel (c): Including bioclimate variables.* Each node is hardwired to another indicator node (green hexagon) that represent the state of various bioclimate variables. *Panel (d): Allowing for missing data.* The model can be further improved by connecting all nodes to a latent node that represents unobserved effects. The observation status at a node is, in the first instance, predicted by the spatial neighbourhood, the bioclimate variables and/or the latent variable. Only if the explanatory power of these correction schemes is not sufficient will there be an incentive for the inference scheme to include further edges related to species interactions. Hence the effect of these corrections is to reduce the network connectivity and filter out spurious interactions.

at a node is, in the first instance, predicted by the spatial neighbourhood. Only if the explanatory power of the latter is not sufficient will there be an incentive for the inference scheme to include further edges related to species interactions. The application to regression based approaches is discussed in [DH5].

Additionally, we want to include bio-climate covariates. In the study described in [DH5], they were related to the average temperature and water availability. These extra variables are included in the same way as the spatial autocorrelation variable. In particular, in a Bayesian network, we introduce fixed connections between the bio-climate covariates and the other nodes.[4] An illustration is given in Figure 8.2(c).

[4]Note that we have to modify the fan-in restriction so that it does not take these extra variables into account. For instance, if the fan-in limit is three, then that means that a species can have up to six parent nodes: three other species, the covariates, and the spatial autocorrelation node.

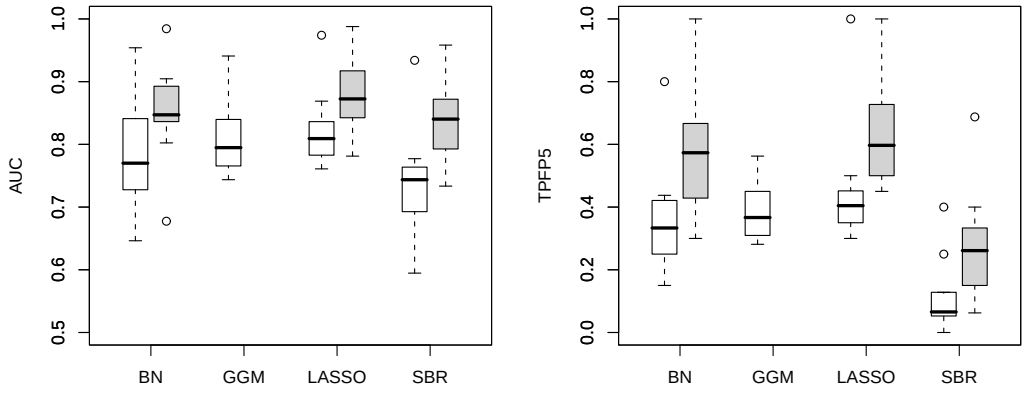


Figure 8.3: **Network reconstruction accuracy.** The figure shows AUC and TPFP5 performance measures for the simulated species population densities. These measures were explained in Section 6.3.2. AUC is the area under the ROC curve. TPFP5 is the true prediction rate at a fixed false positive rate of 5%. The boxplots show the distributions of these scores, where the horizontal bar shows the median, the box margins show the 25th and 75th percentiles, the whiskers indicate data within 2 times the interquartile range, and circles are outliers. Shaded boxes represent models which include spatial autocorrelation. The expected random performance scores are AUC=0.5 and TPFP5=0.05. Four reconstruction methods were compared. BN: Bayesian networks; GGM: graphical Gaussian models; LASSO: L1-penalized regression; and SBR: sparse Bayesian regression. A detailed description of these methods can be found in [DH5].

8.3 Findings

In order to have an objective measure of network recovery, we first tested the ability of the methods to reconstruct the true network structure from test data generated by an ecological simulation model. This model combines a niche model [153] with a stochastic population model [79, Chapter 8] in a 2-dimensional lattice and includes predator-prey interactions in the Lotka-Volterra form. The system is described by a set of coupled stochastic differential equations, which are numerically integrated up to a fixed point in time. At this stage, species population profiles on the 2-dimensional lattice are recorded and provide the training data for the network reconstruction methods. An example for a selected species is shown in Figure 8.1. A detailed description is available from [DH5].

The results are shown in Figure 8.3. It is evident that allowing for spatial autocorrelation significantly boosts the performance. Bayesian networks and LASSO with autorcorrelation outperform the other models. However, there is no significant difference between Bayesian networks and LASSO with autocorrelation; see also Figure 8.4. Note, though, that the approach of Bayesian learning of Bayesian networks was applied in textbook form, as described e.g. in Chapter 2 of [DH1], and did not include the methodological improvements discussed in Sections 6.3-6.5.

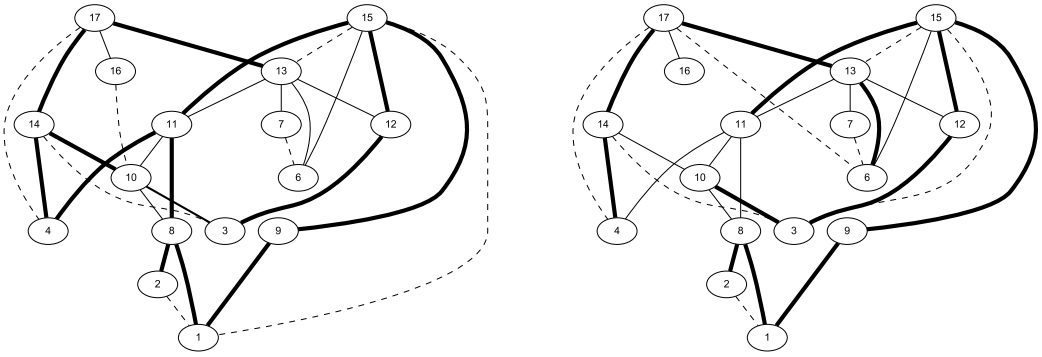


Figure 8.4: **Examples of recovered species interaction networks.** Thick lines represent interactions that were identified correctly (true positives), thin lines represent interactions that were not found (false negatives) and dashed lines are spurious interactions (false positives). The threshold was chosen such that the false positive rate was constant at 5%, resulting in 7 false positive edges. The results from two reconstruction methods are shown. *Left panel*: Reconstruction with Bayesian networks. *Right panel*: Reconstruction with L1-penalized regression (LASSO).

Recovered network	Number of edges	Correlation with phylogenetic distance
Basic data set	80	-0.11 (-0.18,-0.04)
Spatial autocorrelation	69	-0.12 (-0.19,-0.05)
Spatial autocorrelation and bioclimate variables	37	-0.14 (-0.21,-0.07)

Table 8.2: **Results on the European bird atlas data.** The table shows results obtained with Bayesian networks on the 39 warblers from the European bird atlas. The first column indicates the data from which the interaction network was reconstructed. The 'basic data set' includes only the breeding records. The other two data sets also include spatial autocorrelation effects and bioclimate variables. The second column shows the number of interactions whose posterior probability exceeds a given threshold, set to control the number of false positives as described in [DH5]. The third column shows the Pearson correlation between the inferred posterior probability of species interaction and their phylogenetic distance, with the 95% confidence interval shown in brackets.

To test the utility of the available methods for network recovery in large real-world contexts, we applied them to a subset of the European breeding bird data set [55] covering Europe west of 30°E and including all probable and confirmed breeding records. From this data set we extracted the distributions of all 39 old world warbler species breeding in this area. These species are all small insectivores occupying a range of habitat types from boreal forest to Mediterranean reedbeds, several of which are likely to interact at a range of spatial scales. As covariates we included the mean temperature of the coldest month and the water availability for plant growth, two climate variables that had strongest influence on avian distribution [10].

Owing to the lack of a gold standard, the reconstructed network is difficult to assess. A variety of evaluation procedures are discussed in [DH5]. One of them is based on a comparison with phylogenetic results, of the form discussed in Part 1 of this thesis. The conjecture is that species

that are closely related phylogenetically are more likely to show some kind of interaction. The results are summarized in Table 8.2. It turns out that there is indeed a small yet significant anticorrelation between phylogenetic distance and probability of species interaction. This anticorrelation becomes slightly stronger as we include spatial autocorrelation effects and bioclimate variables. Table 8.2 also shows the number of significant interactions. This number decreases as we allow for spatial autocorrelation and include bioclimate variables. These findings suggest that the inclusion of bioclimate variable and the modelling of spatial autocorrelation indeed make the reconstructed network more realistic, in terms of a sightly better correlation with phylogenetic relations, and by pruning out potentially spurious interactions, in corroboration of the concept illustrated in Figure 8.2.

Further methodological advancements aim to allow for unobserved effects, related e.g. to missing species or missing bioclimate variables. This can be modelled by including in the Bayesian network latent nodes that incorporate these effects, as depicted in Figure 8.2(d). The approach we tried in [DH5] was to connect all nodes to a set of binary latent nodes that indicate the presence or absence of some unknown factor. This model is equivalent to connecting all nodes to a single discrete node with an unknown number of discrete activation levels. This model is further equivalent to a mixture model in which the nodes in the network are assigned to the mixture components, the number of which is unknown and has to be inferred from the data. For the practical inference we applied the Bayesian allocation sampler proposed in [DH12]. Unfortunately, owing to the large number of grid locations from which breeding records were taken (over 3000), the MCMC simulations never converged. A future research project is therefore to reduce the computational complexity and replace the free allocation model by a multiple changepoint process, of the form discussed in Sections 4 and 6.5; see also Figure 4.4. The 1-dimensional changepoint processes discussed before have to be generalized to 2-dimensional changepoint processes, to allow for the two spatial coordinates. I am currently pursuing these ideas with Andrej Aderhold and V. Anne Smith.

Part III

My publications

The list below contains a complete list of my publications, as updated in February 2011. In the main text, I cite my papers by number, where the number is enclosed in square brackets and preceded by my initials, DH. Those papers that have been cited in the text are available as PDF files from the following website: http://www.bioss.ac.uk/~dirk/SelectedPublications. The names of the files are identical to the citation labels, except for my book [DH1], which I have divided into chapters. For instance, the twelfth paper from my publication list can be accessed as file DH12.pdf. The second chapter of my book is available in file DH1_chapter2.pdf.

Books

1. **Husmeier D.** , Dybowski R., and Roberts S. (2005)
   Probabilistic Modeling in Bioinformatics and Medical Informatics
   Springer Verlag, New York
2. **Husmeier D.** (1999)
   Neural Networks for Conditional Probability Estimation
   Perspectives in Neural Computation
   Springer Verlag, London

Scientific journals

3. Grzegorczyk M. and **Husmeier D.** (2011)
   Improvements in the reconstruction of time-varying gene regulatory networks: dynamic programming and regularization by information sharing among genes
   Bioinformatics, 27 (5), 693-699.
4. Grzegorczyk M. and **Husmeier D.** (2011)
   Non-homogeneous dynamic Bayesian networks for continuous data
   Machine Learning, in print
5. Faisal A., Dondelinger F., **Husmeier D.** , and Beale C.M. (2010)
   Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods
   Ecological Informatics, 5 (6), 451-464.
6. Grzegorczyk M., **Husmeier D.** and Rahnenführer J. (2010)
   Modelling Non-Stationary Gene Regulatory Processes
   Advances in Bioinformatics, Volume 2010,
   Article ID 749848, doi:10.1155/2010/749848
7. Grzegorczyk M., **Husmeier D.** and Rahnenführer J. (2010)
   Modelling non-stationary dynamic gene regulatory processes with the BGM model
   Computational Statistics, Published online: 15 June 2010

8. Lin K. and **Husmeier D.** (2009)
   Modelling transcriptional regulation with a mixture of factor analyzers and variational Bayesian Expectation Maximization
   EURASIP Journal on Bioinformatics and Systems Biology,
   Article ID 601068, doi:10.1155/2009/601068

9. Lehrach W.P. and **Husmeier D.** (2009)
   Segmenting bacterial and viral DNA sequence alignments with a trans-dimensional phylogenetic factorial hidden Markov model
   Applied Statistics, 58 (3), 307-327

10. **Husmeier D.** and Mantzaris, A.V. (2008)
    Addressing the Shortcomings of Three Recent Bayesian Methods for Detecting Interspecific Recombination in DNA Sequence Alignments
    Statistical Applications in Genetics and Molecular Biology, Vol. 7 : Iss. 1, Article 34.

11. Milne I., Lindner D., Bayer M., **Husmeier D.**, McGuire G., Marshall D.F., and Wright F. (2008)
    TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops
    Bioinformatics, 25(1):126-127.

12. Grzegorczyk M., **Husmeier D.** , Edwards K., Ghazal P., and Millar A. (2008)
    Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler
    Bioinformatics 24(18):2071-2078.

13. Werhli A.V. and **Husmeier D.** (2008)
    Gene Regulatory Network Reconstruction by Bayesian Integration of Prior Knowledge and/or Different Experimental Conditions
    Journal of Bioinformatics and Computational Biology 6 (3), 543-572

14. Grzegorczyk M. and **Husmeier D.** (2008)
    Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move
    Machine Learning 71 (2-3), 265-305.

15. Armstrong M.R., **Husmeier D.** , Phillips M.S. and Bloks V.C. (2007)
    Segregation and recombination of a multipartite mitochondrial DNA in populations of the potato cyst nematode *Globodera pallida*
    Journal of Molecular Evolution 64, 689-701.

16. Werhli A.V. and **Husmeier D.** (2007)
    Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge
    Statistical Applications in Genetics and Molecular Biology, Vol. 6 : Iss. 1, Article 15.

17. Kedzierska A. and **Husmeier D.** (2006)
A Heuristic Bayesian Method for Segmenting DNA Sequence Alignments and Detecting Evidence for Recombination and Gene Conversion
Statistical Applications in Genetics and Molecular Biology, 5 (1), Article 27.

18. Werhli A.V., Grzegorczyk M. and **Husmeier D.** (2006)
Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks
Bioinformatics 22(20): 2523-2531.

19. Lehrach W. P., **Husmeier D.** and Williams C. K. I. (2006)
A regularized discriminative model for the prediction of protein-peptide interactions
Bioinformatics 22: 532-540.

20. **Husmeier D.** (2005)
Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models
Bioinformatics 21: ii166-ii172.

21. **Husmeier D.** , Wright F., Milne I. (2005)
Detecting interspecific recombination with a pruned probabilistic divergence measure
Bioinformatics 21(9):1797-1806

22. Milne I., Wright F., Rowe G., Marshall D.F., **Husmeier D.** , McGuire G. (2004)
TOPALi: software for automatic identification of recombinant sequences within DNA multiple alignments
Bioinformatics 20: 1806-1807

23. **Husmeier D.** (2003)
Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks
Bioinformatics 19: 2271-2282.

24. **Husmeier D.** (2003)
Reverse engineering of genetic networks with Bayesian networks
Biochemical Society Transactions 31 (6): 1516-1518.

25. **Husmeier D.** , McGuire G. (2003)
Detecting Recombination in 4-Taxa DNA Sequence Alignments with Bayesian Hidden Markov Models and Markov Chain Monte Carlo
Molecular Biology and Evolution 20(3):315-337.

26. **Husmeier D.** , McGuire G. (2002)
Detecting recombination with MCMC
Bioinformatics 18: S345-S353.

27. **Husmeier D.** , Wright F. (2002)
A Bayesian Approach to Discriminate between Alternative DNA Sequence Segmentations
Bioinformatics 18 (2), 226-234.

28. **Husmeier D.** , Wright F. (2001)
Detection of Recombination in DNA Multiple Alignments with Hidden Markov Models
Journal of Computational Biology 8 (4), 401-427.

29. **Husmeier D.** , Wright F. (2001)
Probabilistic Divergence Measures for Detecting Interspecies Recombination
Bioinformatics 17, Suppl. 1, S123-S131.

30. Althoefer K., Krekelberg B., **Husmeier D.** , Seneviratne L. (2001)
Reinforcement learning in a rule-based navigator for robotic manipulators
Neurocomputing 37 (1-4), 51-70.

31. **Husmeier D.** (2000)
The Bayesian Evidence Scheme for Regularising Probability-Density Estimating Neural Networks
Neural Computation 12 (11), 2685-2717.

32. **Husmeier D.** (2000): Learning Non-Stationary Conditional Probability Distributions
Neural Networks 13, 287-290.

33. **Husmeier D.** (2000)
Bayesian Regularization of Hidden Markov Models with an Application to Bioinformatics
Neural Network World 10 (4), 589-595.

34. **Husmeier D.** , Penny W., Roberts S.J. (1999)
An Empirical Evaluation of Bayesian Sampling with Hybrid Monte Carlo for Training Neural Network Classifiers
Neural Networks 12, 677-705.

35. Roberts S.J., **Husmeier D.** , Rezek I., Penny W. (1998)
Bayesian Approaches to Gaussian Mixture Modeling
IEEE Transactions on Pattern Analysis and Machine Learning 20 (11), 1133-1142.

36. **Husmeier D.** , Althoefer K. (1998)
Modelling conditional probabilities with network committees: how overfitting can be useful
Neural Network World 8 (4), 417-439.

37. **Husmeier D.** , Taylor J.G. (1998)
Neural Network for Predicting Conditional Probability Densities: Improved Training Scheme Combining EM and RVFL
Neural Networks 11 (1), 89-116.

38. **Husmeier D.** , Taylor J.G. (1997)
Predicting Conditional Probability Densities of Stationary Stochastic Time Series
Neural Networks 10 (3), 479-497.

39. Steinhoff H.J., Schlitter J., Redhardt A., **Husmeier D.** , Zander N. (1992)
Structural fluctuations and conformational entropy in proteins: entropy balance in an intramolecular reaction in methemoglobin
Biochimica et Biophysica Acta 1121, 189-198.

40. Schlitter J., **Husmeier D.** (1992)
System Relaxation and Thermodynamic Integration
Molecular Simulation 8, 285-295.

Book Chapters and Published Proceedings of International Conferences

41. **Husmeier D.**, Dondelinger F., and Lebre S. (2010)
Inter-time segment information sharing for non-homogeneous dynamic Bayesian networks
Advances in Neural Information Processing Systems (NIPS)
42. Dondelinger F., Lebre S., and **Husmeier D.** (2010)
Heterogeneous Continuous Dynamic Bayesian Networks with Flexible Structure and Inter-Time Segment Information Sharing
Proceedings of the International Conference on Machine Learning (ICML), Eds. Furnkranz J., Joachims T., 303-310. Madison, Wisconsin, USA. ISBN 978-1-60558-907-7.
43. Lin K. and **Husmeier D.** (2010)
Mixtures of factor analyzers for modeling transcriptional regulation
In Lawrence, Girolami, Rattray and Sanguinetti (eds.): Learning and Inference in Computational Systems Biology, MIT press, Cambridge, MA, ISBN 9780262013864, pages 153-200.
44. Grzegorczyk M. and **Husmeier D.** (2009)
Non-stationary continuous dynamic Bayesian networks
In Bengio, Schuurmans, Lafferty, Williams and Culotta (eds.): Proceedings of the Twenty-Third Annual Conference on Advances in Neural Information Processing Systems (NIPS), Curran Associates, ISBN 9781605603520, pages 682–690
45. Grzegorczyk M. and **Husmeier D.** (2009)
Avoiding Spurious Feedback Loops in the Reconstruction of Gene Regulatory Networks with Dynamic Bayesian Networks
In: V. Kadirkamanathan et al. (Eds.): Pattern Recognition in Bioinformatics Lecture Notes in Bioinformatics, Springer-Verlag Berlin Heidelberg, pp. 113-124
46. Mantzaris A.V. and **Husmeier D.** (2009)
Distinguishing Regional from Within-Codon Rate Heterogeneity in DNA Sequence Alignments
In: V. Kadirkamanathan et al. (Eds.): Pattern Recognition in Bioinformatics Lecture Notes in Bioinformatics, Springer-Verlag Berlin Heidelberg, pp. 187-198
47. Grzegorczyk M. and **Husmeier D.** (2009)
Modelling non-stationary gene regulatory processes with a non-homogeneous dynamic Bayesian network and the change point process
In: T. Manninen et al. (Eds.): Proceedings of the Sixth International Workshop on Computational Systems Biology, WCSB 2009, Aarhus, Denmark, pp. 51-54, ISBN 978-952-15-2160-7

48. Grzegorczyk M., **Husmeier D.** and Werhli A.V. (2008)
Reverse Engineering Gene Regulatory Networks with Various Machine Learning Methods In: Emmert-Streib F. and Dehmer M. (editors) Analysis of Microarray Data: A Network-Based Approach Wiley-VCH, Weinheim, 2008, pages 101-142. ISBN 978-3-527-31822-3

49. **Husmeier D.** , Werhli A.V. (2007)
"Bayesian Integration of biological prior knowledge into the reconstruction of gene regulatory networks with Bayesian networks" . In Xu Y. and Markstein P. (eds.): Proceedings of the International Conference on Computational Systems Bioinformatics (CSB 2007), Vol. 6, p. 85-95 ISBN 978-1-86094-872-5

50. Lehrach W.P., **Husmeier D.** and Williams C.K.I. (2006)
Probabilistic *in silico* prediction of protein-peptide interactions In: Eskin E., Ideker T., Raphael B. and Workman C. (editors) Systems Biology and Regulatory Genomics Springer Verlag, San Diego, ISBN 978-3-540-48293-2, pages 188-197

51. **Husmeier D.** (2006)
Detecting Mosaic Structures in DNA Sequence Alignments In: Misra JC (editor) Biomathematics: Modelling and Simulation World Scientific, ISBN 981-238-110-4

52. Werhli A.V., Grzegorczyk M., Chiang M.T. and **Husmeier D.** (2006)
Improved Gibbs sampling for detecting mosaic structures in DNA sequence alignments. In: Urfer W. and Turkman M. A. (editors) Statistics in Genomics and Proteomics Centro Internacional de Matematica, Coimbra, Portugal, ISBN: 989-95011-0-7, pages 23-34.

53. **Husmeier D.** , Wright F. (2001)
Approximate Bayesian Discrimination between Alternative DNA Mosaic Structures
In Wingender E., Hofestaedt R., Liebich I. (eds.): 16th German Conference on Bioinformatics (GCB 2001). ISBN 3-00-008114-3, pages 182-184.

54. **Husmeier D.** , Wright F. (2000)
Detecting Sporadic Recombination in DNA Alignments with Hidden Markov Models
In Bornberg-Bauer E., Rost U., Stoye J., Vingron M. (eds.): 15th German Conference on Bioinformatics (GCB 2000) , Logos Verlag Berlin (ISBN 3-89722-498-4), 19-26.

55. Penny W.D., **Husmeier D.** , Roberts S.J. (1999)
The Bayesian Paradigm: Second Generation Neural Computing
In: Lisboa P.J.G., Ifeachor E.C., Srczepaniak A.S. (Ed.), Artificial Neural Networks in Biomedicine, Springer (ISBN: 1-85233-005-8), 11-23.

56. **Husmeier D.** , Roberts S.J. (1999)
Regularisation of RBF-Networks with the Bayesian Evidence Scheme
International Conference on Artificial Neural Networks (ICANN99) , IEE Press, Edinburgh, 533-538.

57. Penny W.D., **Husmeier D.** , Roberts S.J. (1999)
Covariance-based weighting for optimal combination of network predictions
International Conference on Artificial Neural Networks (ICANN99) IEE Press, Edinburgh, 826-831.

58. **Husmeier D.** , Patton G.S., McClure M.O., Harris J.R.W., Roberts S.J.(1999)
Neural Networks for Predicting Kaposi's Sarcoma
International Joint Conference on Neural Networks (IJCNN99) .

59. **Husmeier D.** , Penny W.D., Roberts S.J. (1998)
Empirical Evaluation of Bayesian Sampling for Neural Classifiers
In: Niklasson L., Boden M., Ziemke T. (eds.): International Conference on Artificial Neural Networks - ICANN '98 , Perspectives in Neural Computing, Springer Verlag (ISBN 3-540-76263-9), 323-328.

60. **Husmeier D.** , Taylor J.G. (1997)
Modelling Conditional Probabilities with Committees of RVFL Networks
In: Gerstner W., Germond A., Hasler M., Nicoud J.D. (eds.): International Conference on Artificial Neural Networks - ICANN '97 , Lecture Notes in Computer Science 1327, Springer Verlag (ISBN 3-540-63631-5), 1053-1058.

61. **Husmeier D.** , Taylor J.G. (1997)
Predicting Conditional Probability Densities with the Gaussian Mixture - RVFL Network
In: Smith G.D., Steele N.C., Albrecht R.F. (Eds.): Artificial Neural Networks and Genetic Algorithms , 477-481, Springer Verlag, ISBN 3-211-83087-1.

62. **Husmeier D.** , Allen D., Taylor J.G. (1997)
A Universal Approximator for Learning Conditional Probability Densities
in Ellacott S.W., Mason J.C., Anderson I.J. (eds.): Mathematics of Neural Networks: Models, Algorithms, and Applications , Kluwer Academic Press, Boston (ISBN: 0-7923-9933-1), 198-203.

63. **Husmeier D.** , Taylor J.G. (1996)
A Neural Network Approach to Predicting Noisy Time Series
in Ludermir T.B. (ed.): Annals of the Third Brazilian Symposium on Neural Networks, 221-226, Recife 1996.

Published Comments

64. **Husmeier D.** (2011)
Contribution to the discussion on "Riemann manifold Langevin and Hamiltonian Monte Carlo methods" by Girolami and Calderhead
Journal of the Royal Statistical Society B, in print

65. **Husmeier D.** and Glasbey C. (2007)
Contribution to the discussion on "Model-based clustering for social networks" by Handcock, Raftery and Tantrum
Journal of the Royal Statistical Society A, 170 (4), 340

66. Glasbey C. and **Husmeier D.** (2004)
Contribution to the discussion on "Clustering objects on subsets of attributes" by Friedman

and Meulman
Journal of the Royal Statistical Society B, 66 (4), 840-841

67. **Husmeier D.** (2002)
Contribution to the discussion on statistical modelling and analysis of genetic data
Journal of the Royal Statistical Society B, 64 (4), 751

Technical reports

68. **Husmeier D.**, Wright F. (2001)
Detecting past recombination events in Potato virus Y genomic sequences using statistical methods
Scottish Crop Research Institute, Annual Report 2000/2001, 158-162, ISBN 0 9058 75176

Theses

69. **Husmeier D.** (1997)
"Modelling Conditional Probability Densities with Neural Networks"
PhD thesis, King's College, London 1997.

70. **Husmeier D.** (1994)
Time Series Prediction with Neural Networks.
MSc dissertation, Department of Mathematics, King's College London.

71. **Husmeier D.** (1991)
Numerische Bestimmung des Lösungsmitteleinflusses auf die thermodynamischen Grössen einer intramolekularen Proteinreaktion. (In German. English translation: Numerical estimation of the influence of the solvent on the thermodynamic entities in an intramolecular protein reaction.)
Diplomarbeit, Department of Biophysics, University of Bochum.

Bibliography

[1] A. Ahmed and E. P. Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106:11878–11883, 2009.

[2] M. Arbeitman, E. Furlong, F. Imam, E. Johnson, B. Null, B. Baker, M. Krasnow, M. Scott, R. Davis, and K. White. Gene expression during the life cycle of Drosophila melanogaster. *Science*, 297(5590):2270–2275, 2002.

[3] P. W. Atkins. *Physical Chemistry*. Oxford University Press, Oxford, 3rd edition, 1986.

[4] N. H. Augustin, M. A. Mugglestone, and S. T. Buckland. An autologistic model for the spatial distribution of wildlife. *J. Appl. Ecol.*, 33(2):339–347, 1996.

[5] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36, 1994.

[6] P. Baldi and P. Brunak. *Bioinformatics - The Machine Learning Approach*. MIT Press, Cambridge, MA, 1998.

[7] H. Bandelt and A. W. M. Dress. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution*, 1:242–252, 1992.

[8] N. Barkai and S. Leibler. Circadian clocks limited by noise. *Nature*, 403:267–268, 2000.

[9] M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[10] C. Beale, J. Lennon, and A. Gimona. Opening the climate envelope reveals no macroscale associations with climate in European birds. *Proceedings of the National Academy of Sciences*, 105(39):14908, 2008.

[11] M. A. Beer and S. Tavazoie. Predicting gene expression from sequence. *Cell*, 117:185–98, April 2004.

[12] B. E. Beisner, D. T. Haydon, and K. Cuddington. Alternative stable states in ecology. *Front. Ecol. Environ.*, 1(7):376–382, 2003.

[13] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, New York, 1995. ISBN 0-19-853864-2.

[14] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Singapore, 2006.

[15] N. Blüthgen, F. Menzel, and N. Blüthgen. Measuring specialization in species interaction networks. *BMC ecology*, 6(1):9, 2006.

[16] A.-L. Boulesteix and K. Strimmer. Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. *Theor Biol Med Model*, 2(1):23, 2005.

[17] R. J. Boys and D. A. Henderson. A comparison of reversible jump MCMC algorithms for DNA sequence segmentation using hidden Markov models. *Computer Science and Statistics*, 33:35–49, 2001.

[18] R. J. Boys and D. A. Henderson. A Bayesian approach to DNA sequence segmentation. *Biometrics*, 60:573–588, 2004.

[19] R. J. Boys, D. A. Henderson, and D. J. Wilkinson. Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Applied Statistics*, 49:269–285, 2000.

[20] H. J. Bussemaker, H. Li, and E. D. Siggia. Regulatory element detection using correlation with expression. *Nat Genet*, 27(2):167–71, 2001.

[21] A. S. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, 2000:418–429, 2000.

[22] A. S. Butte and I. S. Kohane. Relevance networks: A first step toward finding genetic regulatory networks within microarray data. In G. Parmigiani, E. S. Garett, R. A. Irizarry, and S. L. Zeger, editors, *The Analysis of Gene Expression Data*, pages 428–446, New York, 2003. Springer.

[23] B. Calderhead, M. Girolami, and N. D. Lawrence. Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. *Neural Information Processing Systems (NIPS)*, 22, 2008.

[24] I. Cantone, L. Marucci, F. Iorio, M. A. Ricci, V. Belcastro, M. Bansal, S. Santini, M. di Bernardo, D. di Bernardo, and M. P. Cosma1. A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, 137:172181, 2009.

[25] G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[26] T. Chen, H. L. He, and G. M. Church. Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing*, 4:29–40, 1999.

[27] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[28] J. E. Cohen, K. Schoenly, K. L. Heong, H. Justo, G. dArida, A. T. Barrion, and J. Litsinger. A food-web approach to evaluating the effect of insecticide spraying on insect pest population-dynamics in a Philippine irrigated rice ecosystem. *J. Appl. Ecol.*, 31:747–763, 1994.

[29] E. M. Conlon, X. S. Liu, J. D. Lieb, and J. S. Liu. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci*, 100(6):3339–44, 2003.

[30] M. R. T. Dale and M. J. Fortin. Spatial autocorrelation and statistical tests in ecology. *Ecoscience*, 9(2):162–167, 2002.

[31] J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. In *ICML '06: Proceedings of the 23rd international conference on Machine Learning*, pages 233–240, New York, NY, USA, 2006. ACM.

[32] H. De Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103, 2002.

[33] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B39(1):1–38, 1977.

[34] P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, 2000.

[35] M. K. Dougherty, J. Muller, D. A. Ritt, M. Zhou, X. Z. Zhou, T. D. Copeland, T. P. Conrads, T. D. Veenstra, K. P. Lu, and D. K. Morrison. Regulation of Raf-1 by direct feedback phosphorylation. *Molecular Cell*, 17:215–224, 2005.

[36] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis. Probabilistic models of proteins and nucleic acids.* Cambridge University Press, Cambridge, UK, 1998.

[37] B. Efron. Bootstrap methods: Another look at the jacknife. *Annals of Statistics*, 7:1–26, 1979.

[38] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95:14863–14868, 1998.

[39] P. Fearnhead. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16:203–213, 2006.

[40] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27:401–440, 1978.

[41] J. Felsenstein. Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics*, 22:521–565, 1988.

[42] J. Felsenstein. Phylip. Free package of programs for inferring phylogenies, available from http://evolution.genetics.washington.edu/phylip.html, 1996.

[43] J. Felsenstein. The troubled growth of statistical phylogenetics. *Systems Biology*, 50(4):465–467, 2001.

[44] W. M. Fitch. Towards defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20:406–416, 1971.

[45] N. Friedman and D. Koller. Being Bayesian about network structure. *Machine Learning*, 50:95–126, 2003.

[46] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000.

[47] T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, 403:339–342, 2000.

[48] D. Geiger and D. Heckerman. Learning Gaussian networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 235–243, San Francisco, CA., 1994. Morgan Kaufmann.

[49] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472, 1992.

[50] Z. Ghahramani and G. E. Hinton. The EM algorithm for mixtures of factor analyzers. Technical report, CRG-TR-96-1, University of Toronto, 1996.

[51] P. Giudici and R. Castelo. Improving Markov chain Monte Carlo model search for data mining. *Machine Learning*, 50:127–158, 2003.

[52] N. Goldman. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. *Systematic Zoology*, 39:345–361, 1990.

[53] N. C. Grassly and E. C. Holmes. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Molecular Biology and Evolution*, 14(3):239–247, 1997.

[54] P. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[55] W. J. M. Hagemeijer and M. J. Blair. *The EBCC atlas of European breeding birds: their distribution and abundance*. Poyser London, 1997.

[56] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok,

M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, Sept. 2004.

[57] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pacific Symposium on Biocomputing*, 6:422–433, 2001.

[58] M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.

[59] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[60] D. Heckerman. A tutorial on learning with Bayesian networks. In M. I. Jordan, editor, *Learning in Graphical Models*, Adaptive Computation and Machine Learning, pages 301–354, Cambridge, Massachusetts, 1999. MIT Press.

[61] D. Heckerman and D. Geiger. Learning Bayesian networks: A unification for discrete and Gaussian domains. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 274–82, San Francisco, CA, 1995. Morgan Kaufmann.

[62] D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:245–274, 1995.

[63] J. Hein. A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, 36:396–405, 1993.

[64] M. L. Henneman and J. Memmott. Infiltration of a Hawaiian community by introduced biological control agents. *Science*, 293(5533):1314–1316, 2001.

[65] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation*. Addison Wesley, Redwood City, CA, 1991.

[66] P. G. Hoel. *Introduction to Mathematical Statistics*. John Wiley and Sons, Singapore, 1984.

[67] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology*, 296(5):1205–14, March 2000.

[68] S. Imoto, T. Higuchi, T. Goto, S. Kuhara, and S. Miyano. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Proceedings IEEE Computer Society Bioinformatics Conference*, (CSB'03):104–113, 2003.

[69] T. C. Ings, J. M. Montoya, J. Bascompte, N. Bluthgen, L. Brown, C. F. Dormann, F. Edwards, D. Figueroa, U. Jacob, J. I. Jones, R. B. Lauridsen, M. E. Ledger, H. M. Lewis, J. M. Olesen, F. J. F. van Veen, P. H. Warren, and G. Woodward. Review: Ecological networks beyond food webs. *J. Anim. Ecol.*, 78:253–269, 2009.

[70] H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23:187–200, 1958.

[71] M. Kanehisa. A database for post-genome analysis. *Trends Genet*, 13:375–376, 1997.

[72] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28:27–30, 2000.

[73] M. Kanehisa, S. Goto, M. Hattori, K. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34:D354–357, 2006.

[74] K. C. Kao, Y.-L. Yang, R. Boscolo, C. Sabatti, V. Roychowdhury, and J. C. Liao. Transcriptome-based determination of multiple transcription regulator activities in *Escherichia coli* by using network component analysis. *Proc Natl Acad Sci*, 101(2):641–6, 2004.

[75] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[76] Y. Ko, C. Zhai, and S. Rodriguez-Zas. Inference of gene pathways using Gaussian mixture models. In *BIBM International Conference on Bioinformatics and Biomedicine*, pages 362–367. Fremont, CA, 2007.

[77] P. J. Krause. Learning probabilistic networks. *Knowledge Engineering Review*, 13:321–351, 1998.

[78] W. J. Krzanowski and F. H. C. Marriott. *Multivariate Analysis*, volume 2. Arnold, 1995. ISBN 0-340-59325-3.

[79] R. Lande, S. Engen, and B. Saether. *Stochastic Population Dynamics in Ecology and Conservation.* Oxford University Press, 2003.

[80] B. Larget and D. L. Simon. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16(6):750–759, 1999.

[81] C. Lawrence, S. Altschul, M. Boguski, J. Liu, A. Neuwald, and W. JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

[82] N. D. Lawrence and M. Rattray. A brief introduction to Bayesian inference. In R. Lawrence, Girolami and Sanguinetti, editors, *Learning and Inference in Computational Systems Biology*, Computational Molecular Biology, pages 95–114. MIT Press, 2010.

[83] S. Lèbre. *Stochastic process analysis for Genomics and Dynamic Bayesian Networks inference.* PhD thesis, Université d'Evry-Val-d'Essonne, France, 2007.

[84] S. Lèbre, J. Becq, F. Devaux, G. Lelandais, and M. Stumpf. Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology*, 4(130), 2010.

[85] J. J. Lennon. Red-shifts and red herrings in geographical ecology. *Ecography*, 23:101–113, 2000.

[86] J. C. Liao, R. Boscolo, Y.-L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury. Network component analysis: Reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci*, 100(26):15522–7, 2003.

[87] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.

[88] J. S. Liu, A. F. Neuwald, and C. E. Lawrence. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, 90(432):1156–1170, 1995.

[89] D. Maddison. The discovery and importance of multiple islands of most parsimonious trees. *Systematic Zoology*, 40:315–328, 1991.

[90] D. Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232, 1995.

[91] A. V. Mantzaris. Detecting mosaic structures in dna sequence alignments. Master's thesis, School of Informatics, University of Edinburgh, 2006.

[92] J. Maynard Smith. Analyzing the mosaic structure of genes. *Journal of Molecular Evolution*, 34:126–129, 1992.

[93] G. McGuire and F. Wright. TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics*, 16(2):130–134, 2000.

[94] G. McGuire, F. Wright, and M. Prentice. A graphical method for detecting recombination in phylogenetic data sets. *Molecular Biology and Evolution*, 14(11):1125–1131, 1997.

[95] G. McGuire, F. Wright, and M. Prentice. A Bayesian method for detecting past recombination events in DNA multiple alignments. *Journal of Computational Biology*, 7(1/2):159–170, 2000.

[96] J. Memmott. The structure of a plant-pollinator food web. *Ecology Letters*, 2(5):276–280, 1999.

[97] J. Memmott, S. Fowler, Q. Paynter, A. Sheppard, and P. Syrett. The invertebrate fauna on broom, *Cytisus scoparius*, in two native and two exotic habitats. *Acta Oecol.*, 21(3):213–222, 2000.

[98] M. Middendorf, A. Kundaje, M. Shah, Y. Freund, C. Wiggins, and C. Leslie. Motif discovery through predictive modeling of gene regulation. In *Proceedings of RECOMB 2005*, pages 538–52, 2005.

[99] M. Middendorf, A. Kundaje, C. Wiggins, Y. Freund, and C. Leslie. Predicting genetic regulatory response using classification. *Bioinformatics*, 20(Suppl 1):i232–40, 2004.

[100] V. N. Minin, K. S. Dorman, F. Fang, and M. A. Suchard. Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, 21(13):3034–3042, 2005.

[101] S. Mnaimneh, A. P. Davierwala, J. Haynes, J. Moffat, W.-T. Peng, W. Zhang, X. Yang, J. Pootoolal, G. Chua, A. Lopez, M. Trochesset, D. Morse, N. J. Krogan, S. L. Hiley, Z. Li, Q. Morris, J. Grigull, N. Mitsakakis, C. J. Roberts, J. F. Greenblatt, C. Boone, C. A. Kaiser, B. J. Andrews, and T. R. Hughes. Exploration of essential gene functions via titratable promoter alleles. *Cell*, 118(1):31–44, Jul 2004.

[102] A. Nobile and A. Fearnside. Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, 17(2):147–162, 2007.

[103] R. D. M. Page and E. C. Holmes. *Molecular Evolution - A Phylogenetic Approach*. Blackwell Science, Cambridge (UK), 1998.

[104] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Singapore, 3rd edition, 1991.

[105] T. Pawson and J. D. Scott. Signaling Through Scaffold, Anchoring, and Adaptor Proteins. *Science*, 278(5346):2075–2080, 1997.

[106] T. M. Phuong, D. Lee, and K. H. Lee. Regression trees for regulatory element identification. *Bioinformatics*, 20(5):750–7, 2004.

[107] I. Pournara and L. Wernisch. Reconstruction of gene networks using Bayesian learning and manipulation experiments. *Bioinformatics*, 20:2934–2942, 2004.

[108] I. Pournara and L. Wernisch. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, 8:61, 2007.

[109] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[110] S. Raychaudhuri, J. M. Stuart, and R. B. Altman. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput*, 5:455–66, 2000.

[111] D. J. Reiss and B. Schwikowski. Predicting protein-peptide interactions via a network-based motif sampler. *Bioinformatics*, 20(suppl1):i274–282, 2004.

[112] A. Remenyi, H. R. Scholer, and M. Wilmanns. Combinatorial control of gene expression. *Nat Struct Mol Biol*, 11(9):812–5, 2004.

[113] J. J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[114] C. P. Robert, G. Celeux, and J. Diebolt. Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statistics & Probability Letters*, 16:77–83, 1993.

[115] C. P. Robert, T. Ryden, and D. M. Titterington. Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society, Series B*, 62(1):57–75, 2000.

[116] D. L. Robertson, P. M. Sharp, F. E. McCutchan, and B. H. Hahn. Recombination in HIV-1. *Nature*, 374:124–126, 1995.

[117] D. Robinson and L. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

[118] J. W. Robinson and A. J. Hartemink. Non-stationary dynamic Bayesian networks. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 21, pages 1369–1376. Morgan Kaufmann Publishers, 2009.

[119] S. Rogers and M. Girolami. A Bayesian regression approach to the inference of regulatory networks from gene expression data. *Bioinformatics*, 21(14):3131–3137, 2005.

[120] J. Ruan and W. Zhang. A bi-dimensional regression tree approach to the modeling of gene expression regulation. *Bioinformatics*, 22(3):332–40, 2006.

[121] R. Y. Rubinstein. *Simulation and the Monte Carlo Method*. John Wiley & Sons, New York, 1981.

[122] C. Sabatti and G. M. James. Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22(6):739–46, 2006.

[123] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Protein-signaling networks derived from multiparameter single-cell data. *Science*, 308:523–529, 2005.

[124] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

[125] G. Sanguinetti, M. Rattray, and N. D. Lawrence. A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription. *Bioinformatics*, 22(14):1753–9, 2006.

[126] J. Schäfer and K. Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

[127] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 32, 2005.

[128] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[129] E. Segal, Y. Barash, I. Simon, N. Friedman, and D. Koller. From promoter sequence to expression: A probabilistic framework. In *RECOMB 2002 Conference Proceedings*, 2002.

[130] E. Segal and R. Sharan. A discriminative model for identifying spatial cis-regulatory mdoules. In *RECOMB 2004 Conference Proceedings*, 2004.

[131] E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, 19(Suppl.1):i272–i282, 2003.

[132] Y. Shi, I. Simon, T. Mitchell, and Z. Bar-Joseph. A combined expression-interaction model for inferring the temporal activity of transcription factors. In M. Vingron and L. Wong, editors, *RECOMB: 12th Annual International Conference on Research in Computational Molecular Biology*, volume 4955 of *Lecture Notes in Computer Science*, pages 82–97, Singapore, 2008. Springer.

[133] A. R. E. Sinclair and A. E. Byrom. Understanding ecosystem dynamics for conservation of biota. *J. Anim. Ecol.*, 75:64–79, 2006.

[134] M. A. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9:91–116, 1992.

[135] C. Stockham, L.-S. Wang, and T. Warnow. Statistically based postprocessing of phylogenetic analysis by clustering. *Bioinformatics*, 18:S285–S293, 2002.

[136] K. Strimmer and V. Moulton. Likelihood analysis of phylogenetic networks using directed graphical models. *Molecular Biology and Evolution*, 17(6):875–881, 2000.

[137] K. Strimmer, C. Wiuf, and V. Moulton. Recombination analysis using directed graphical models. *Molecular Biology and Evolution*, 18(1):97–99, 2001.

[138] J. A. Studier and K. J. Keppler. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution*, 5:729–731, 1988.

[139] M. A. Suchard, R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer. Inferring spatial phylogenetic variation along nucleotide sequences: A multiple changepoint model. *Journal of the American Statistical Association*, 98(462):427–437, 2003.

[140] M. Sudol and T. Hunter. New wrinkles for an old domain. *Cell*, 103:1001–1004, 2000.

[141] M. C. Teixeira, P. Monteiro, P. Jain, S. Tenreiro, A. R. Fernandes, N. P. Mira, M. Alenquer, A. T. Freitas, A. L. Oliveira, and I. Sa-Correia. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 34(Database issue):446–451, Jan 2006.

[142] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[143] M. Tipping and A. Faul. Fast marginal likelihood maximisation for sparse Bayesian models. In B. C. M. and F. B. J., editors, *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 9, 2003.

[144] A. H. Y. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. V. Hogue, S. Fields, C. Boone, and G. Cesareni. A Combined Experimental and Computational Strategy to Define Protein Interaction Networks for Peptide Recognition Modules. *Science*, 295(5553):321–324, 2002.

[145] C. Tuffley and M. Steel. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, 59:581–607, 1997.

[146] R. Twyman. *Principles of Proteomics*. BIOS Scientific Publishers, New York, 2004.

[147] E. P. van Someren, B. L. T. Vaes, W. T. Steegenga, A. M. Sijbers, K. J. Dechering, and M. J. T. Reinders. Least absolute regression network analysis of the murine osterblast differentiation network. *Bioinformatics*, 22(4):477–484, 2006.

[148] D. P. Vázquez and D. Simberloff. Ecological specialization and susceptibility to disturbance: Conjectures and refutations. *The American Naturalist*, 159(6):606–623, June 2002.

[149] V. Vyshemirsky and M. A. Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839, 2008.

[150] A. V. Werhli. *Reconstruction of gene regulatory networks from postgenomic data*. PhD thesis, Biomathematics & Statistics Scotland (BioSS) and University of Edinburgh, 2007.

[151] D. Wilkinson. *Stochastic Modelling for Systems Biology*. Chapman & Hall, 2006.

[152] P. M. Williams. Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7:117–143, 1995.

[153] R. Williams and N. Martinez. Simple rules yield complex food webs. *Nature*, 404(6774):180–183, 2000.

[154] X. Yu, J. Lin, D. J. Zack, and J. Qian. Identification of tissue-specific cis-regulatory modules based on interactions between transcription factors. *BMC Bioinformatics*, 8, 2007.

[155] D. E. Zak, F. J. Doyle, G. E. Gonye, and J. S. Schwaber. Simulation studies for the identification of genetic networks from cDNA array and regulatory activity data. In *Proceedings of the Second International Conference on Systems Biology*, pages 231–238, Pasadena, CA., November 2001.

[156] D. E. Zak, F. J. Doyle, and J. S. Schwaber. Local identifiability: when can genetic networks be identified from microarray data? *Proceedings of the Third International Conference on Systems Biology*, pages 236–237, 2002.

[157] J. Zhou and B. G. Spratt. Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. *Molecular Microbiology*, 6:2135–2146, 1992.