CHAPTER 1

DETECTING MOSAIC STRUCTURES IN DNA SEQUENCE ALIGNMENTS

DIRK HUSMEIER

This article first provides a concise introduction to the statistical approach to phylogenetics. It then describes a new method for detecting mosaic structures in DNA sequence alignments, which is based on combining two probabilistic graphical models: (1) a taxon graph (phylogenetic tree) representing the relationships among the taxa, and (2) a site graph (hidden Markov model) representing spatial correlations between nucleotides.

1. Introduction

The recent advent of multiple-resistant pathogens has led to an increased interest in interspecies 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 [25]. 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 [16]. Thus interspecies recombination, illustrated in Figs. 8 and 9, raises the possibility that bacteria and viruses can acquire biologically important traits through the exchange and transfer of genetic material.

In the last few years, a plethora of methods for detecting interspecies recombination have been developed — following up on the seminal paper by John Maynard Smith [16] — and it is beyond the scope of this article to provide a comprehensive overview. Instead, the focus will be on a novel approach, in which two probabilistic graphical models are combined: (1) a taxon graph (phylogenetic tree) representing the relationships among the species or strains, and (2) a site graph (hidden Markov model) representing

interactions between different sites in the DNA sequence alignments. While at present this approach is still limited to deal with only small numbers of species or strains simultaneously, it has two advantages over existing (mostly heuristic) methods: first, it can predict the locations and breakpoints of recombinant regions more accurately than what can be achieved with most existing techniques. Second, it provides a proper probabilistic generative model. This implies that well-known methods from statistics, like maximum likelihood, can be applied to estimate the parameters. It also renders the model amenable to established statistical methods of hypothesis testing and model selection.

The article is organized as follows. Section 2 provides a brief introduction to the statistical approach to phylogenetics. Section 3 explains the biological process of interspecific recombination. Section 4 provides a short recapitulation of hidden Markov models. Section 5 discusses a hybrid model — combining phylogenetic trees with hidden Markov models — for detecting recombination in DNA sequence alignments. Also, different ways of estimating the model parameters are discussed. Section 6 describes several DNA sequence alignments, on which the proposed model and training algorithms are tested. The results of these tests are discussed in Sec. 7. Finally, Sec. 8 contains a short summary and an outlook on future work.

2. A Brief Introduction to Phylogenetics

2.1. *Topology and parameters of a phylogenetic tree*

The objective of phylogenetics is to infer the evolutionary relationships among different species or strains and to display them in a tree-structured graphical model called a *phylogenetic tree*. An example is given in Fig. 1. The leaves of the (unrooted) 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 example, warm-blooded from cold-blooded animals, birds from mammals, primates from rodents, and so on. 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 Fig. 1 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

Frog	G C T T G A C T T C T G A G G T T
Chicken	G C G T A A C T T C A C A T G A T
Human	G C G T C A C T T G A G A C G C T
Rabbit	G C G T C A C T T G A G A C G C T
Mouse	G C G T C A C T T G A C A G G C T
Opossum	G C G T C A C T T G A G A C G C T

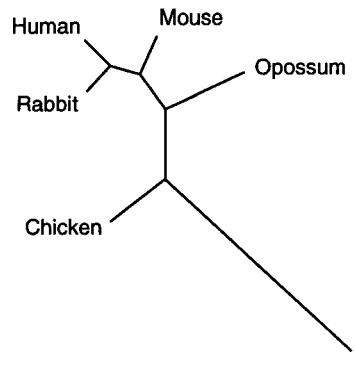

Fig. 1. **Phylogeny and DNA sequence alignment.** The figure shows a phylogenetic tree for six species and a subregion of the DNA sequence alignment from which it is inferred. The *topology* of the tree is the branching order, that is, the way the species are distributed across the leaf nodes. The *parameters* of the tree are the branch lengths, which represent phylogenetic time.

former group, opossum is grouped out, since it is a marsupial and therefore 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, $(2n-5)!!$ different (unrooted) tree topologies, as can easily be proved by induction (see, for instance [4, Chap. 7]). In what follows, we will use the integer variable $S \in \{1, 2, \ldots, (2n-5)!!\}$ to label the different tree topologies.

The second type of information we obtain from a phylogenetic tree are the *branch lengths*, which represent phylogenetic time, measured by the average amount of mutational change. For example, Fig. 1 shows a comparatively long branch leading to the leaf with *frog*. This suggests that the splitting of the lineages separating *frog* from the other animals happened comparatively long ago, that is, earlier than the other speciation events. This is a reasonable conjecture as *frog* is the only cold-blooded animal, whereas all the other animals are warm-blooded. A (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 $\mathbf{w} = (w_1, \ldots, w_{2n-3})$.

2.2. *DNA sequences and sequence alignments*

We now need a method to infer the correct topology of a tree and its branch lengths for a given set of species. As the driving force for evolution are mutations, that is, errors in the replication of DNA, it is reasonable to base our inference process on this information. This approach has recently become viable by major breakthroughs in DNA sequencing techniques. In July 1995, the entire 1.8 million base pairs of the genome of *Haemophilus influenzae*, a small Gram-negative bacterium, was published. Since then, the amount of DNA sequence data in publicly accessible data bases has been growing exponentially and is now about to claim its biggest triumph: the complete 3.3 billion base-pair DNA sequence of the entire human genome (for which a first draft was already released in June 2000).

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 want to compare homologous subsequences, that is, regions of the genome that have been acquired from the same common ancestor. Also, one has to allow for nucleotide insertions and deletions. For example, a direct comparison of the sequences

```
A C G T T A T A
A G T C A T A
```

gives the erroneously small count of only a single site with identical nucleotides. This is due to the insertion of a C in the second position of the first strand, or, equivalently, the deletion of a nucleotide at the second position of the second strand (the insertion of a so-called *gap*). A correct comparison leads to

```
A C G T T A T A
A - G T C A T A
```

which suggests that the sequences differ in only two positions. The process of (1) finding homologous DNA subsequences and (2) correcting for

insertions and deletions is called DNA sequence alignment. A standard algorithm is Clustal-W, discussed in [28]. The details are beyond the scope of this article.

Figure 1, top, shows a small section of the DNA sequence alignment used for inferring the tree at the bottom of Fig. 1. Rows represent different species or strains (generic name: taxa), columns represent different sites or positions on the DNA. At the majority of sites, all nucleotides are identical, which reflects the fact that the compared sequences 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. Note, however, that the process of nucleotide substitution is intrinsically stochastic. We will therefore discuss, in the following two subsections, a mathematical model for statistical phylogenetic inference.

2.3. *A mathematical model of nucleotide substitution*

The driving force for evolution are nucleotide substitutions, which can be modelled as transitions in a 4-state state space, shown in Fig. 2. $P(Y|X,w)$, 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 mutation rate λ with physical time t: $w = \lambda t$. 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 Fig. 2. 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 $P(T|A,w)$. This increase is faster for a mutation within a nucleotide class (purine $\rightarrow$ purine, pyrimidine $\rightarrow$ pyrimidine), than between nucleotide classes (purine $\leftrightarrow$ pyrimidine). For $w \rightarrow \infty$, the system "forgets" its initial configuration as the result of the mixing caused by an increasing number of nucleotide substitutions. Consequently, $P(Y|X,w) \rightarrow \Pi(Y)$,

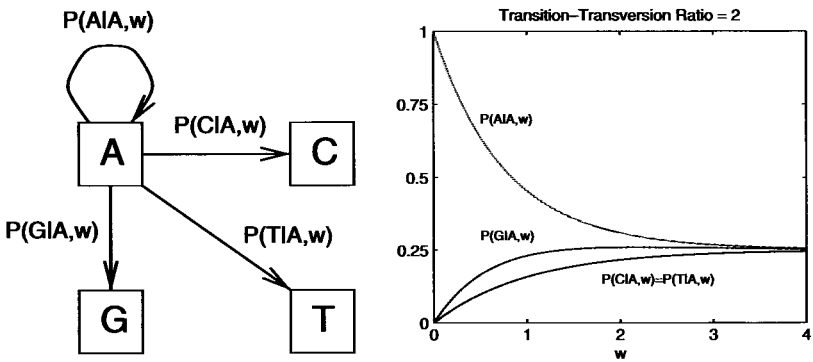

Fig. 2. **Mathematical model of nucleotide substitutions.** *Left:* Nucleotide substitutions are modelled as transitions in a 4-state state space. The transition probabilities depend on the phylogenetic time $w = \alpha t$, where t is physical time and α is a mutation rate. *Right:* Dependence of the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

where $X, Y \in \{A, C, G, T\}$, and $\Pi(Y)$ is the equilibrium distribution (here $\Pi(Y) = 1/4 \,\forall\, Y$).

Let $y_i(t) \in \{A, C, G, T\}$ denote the nucleotide at site i and at physical time t. This notation will be used throughout this chapter: the subscript refers to the position in the alignment, while the expression in brackets denotes physical or (later) phylogenetic time. The total length of the alignment is N, that is, $i \in \{1, \ldots, N\}$. The derivation of the aforementioned results is based on the theory of homogeneous Markov chains and the following assumptions:

- The process is Markov:

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

- The Markov process is homogeneous:

$$P\big(y_i(s+t)|y_i(s)\big) = P\big(y_i(t)|y_i(0)\big).$$

- The Markov process is the same for all positions:

$$P\big(y_i(t)|y_i(0)\big) = P\big(y_k(t)|y_k(0)\big) \quad \forall i, k \in \{1, \ldots, N\}.$$

- Substitutions at different positions are independent of each other:

$$P\big(y_1(t), \ldots, y_N(t)|y_1(0), \ldots, y_N(0)\big) = \prod_{i=1}^{N} P\big(y_i(t)|y_i(0)\big).$$

This implies that the nucleotide substitution process at a given site is completely specified by the following 4-by-4 transition matrix:

$$\mathbf{P}(t) = \begin{bmatrix} P(y(t)=A|y(0)=A) & \cdots & P(y(t)=A|y(0)=T) \\ P(y(t)=G|y(0)=A) & \cdots & P(y(t)=G|y(0)=T) \\ P(y(t)=C|y(0)=A) & \cdots & P(y(t)=C|y(0)=T) \\ P(y(t)=T|y(0)=A) & \cdots & P(y(t)=T|y(0)=T) \end{bmatrix}. \tag{1}$$

Because of the site independence, the site label (that is, the subscript) has been dropped to simplify the notation. Equation (1) obviously implies that

$$\mathbf{P}(0) = \mathbf{I}, \tag{2}$$

where $\mathbf{I}$ is the unit matrix. We now make the ansatz

$$\mathbf{P}(dt) = \mathbf{P}(0) + \mathbf{R}dt, \tag{3}$$

where $\mathbf{R}$ is the so-called rate matrix. From the theory of homogeneous Markov chains it is known that

$$\mathbf{P}(t+dt) = \mathbf{P}(dt)\mathbf{P}(t), \tag{4}$$

which follows from the Chapman–Kolmogorov equation; see [10] or [22]. Inserting Eqs. (2) and (3) into (4) gives:

$$\mathbf{P}(t+dt) = (\mathbf{I} + \mathbf{R}dt)\mathbf{P}(t) \tag{5}$$

and

$$\frac{d\mathbf{P}}{dt} = \mathbf{R}\mathbf{P}. \tag{6}$$

This is a system of linear differential equations with the solution

$$\mathbf{P}(t) = e^{\mathbf{R}t}. \tag{7}$$

To make sure that $\mathbf{P}(t)$ is a proper transition matrix, that is, has columns that sum to 1, the columns of the rate matrix $\mathbf{R}$ have to sum to 0. A possible design for $\mathbf{R}$, the so-called Kimura model [15], is of the form

$$\mathbf{R} = \begin{bmatrix} -2\beta-\alpha & \beta & \alpha & \beta \\ \beta & -2\beta-\alpha & \beta & \alpha \\ \alpha & \beta & -2\beta-\alpha & \beta \\ \beta & \alpha & \beta & -2\beta-\alpha \end{bmatrix}. \tag{8}$$

Here, the rows (from top to bottom) and columns (from left to right) correspond to the nucleotides A, C, G, T (in the indicated order). The positive parameters α and β denote the rates of transitions (mutations within a

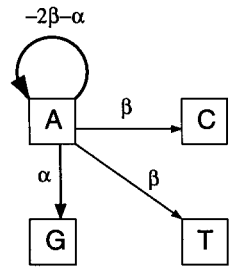

Fig. 3. **Kimura model of nucleotide substitutions.** The figure presents a partial graphical display of the rate matrix of Eq. (8), showing mutations out of nucleotide A. The positive parameter α denotes the rate of a transition (purine $\rightarrow$ purine, pyrimidine $\rightarrow$ pyrimidine), while β denotes the rate of a transversion (purine $\leftrightarrow$ pyrimidine).

nucleotide class: purine $\rightarrow$ purine, pyrimidine $\rightarrow$ pyrimidine) and transversions (mutations between nucleotide classes: purine $\leftrightarrow$ pyrimidine), respectively.[a] An illustration is given in Fig. 3.

It can now be shown [15] that inserting (8) into (7) leads to

$$\mathbf{P}(t) = e^{\mathbf{R}t} = \begin{bmatrix} d(t) & f(t) & g(t) & f(t) \\ f(t) & d(t) & f(t) & g(t) \\ g(t) & f(t) & d(t) & f(t) \\ f(t) & g(t) & f(t) & d(t) \end{bmatrix}, \tag{9}$$

where

$$\begin{aligned} f(t) &= \frac{1}{4}(1 - e^{-4\beta t}), \\ g(t) &= \frac{1}{4}(1 + e^{-4\beta t} - 2e^{-2(\alpha+\beta)t}), \\ d(t) &= 1 - 2f(t) - g(t). \end{aligned}$$

Defining $\lambda = 4\beta$, which implies that the phylogenetic time is given by

$$w = 4\beta t, \tag{10}$$

this results in

$$f(w) = \frac{1}{4}(1 - e^{-w}) \tag{11}$$

[a]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*, 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.

$$g(w) = \frac{1}{4}(1 + e^{-w} - 2e^{-\frac{\tau+1}{2}w}) \tag{12}$$

$$d(w) = 1 - 2f(w) - g(w) \tag{13}$$

in which τ denotes the transition-transversion ratio:

$$\tau = \frac{\alpha}{\beta}. \tag{14}$$

Denoting by $P(Y|X,w)$ the probability that at a given site in the alignment nucleotide Y is observed given that nucleotide X was at this site w phylogenetic time units before, we can re-write $\mathbf{P}$, the transition matrix of (1), as follows:

$$\begin{aligned}\mathbf{P}(w) &= \begin{bmatrix} P(A|A,w) & P(A|C,w) & P(A|G,w) & P(A|T,w) \\ P(G|A,w) & P(G|C,w) & P(G|G,w) & P(G|T,w) \\ P(C|A,w) & P(C|C,w) & P(C|G,w) & P(C|T,w) \\ P(T|A,w) & P(T|C,w) & P(T|G,w) & P(T|T,w) \end{bmatrix} \\ &= \begin{bmatrix} d(w) & f(w) & g(w) & f(w) \\ f(w) & d(w) & f(w) & g(w) \\ g(w) & f(w) & d(w) & f(w) \\ f(w) & g(w) & f(w) & d(w) \end{bmatrix}, \end{aligned} \tag{15}$$

where $d(w)$, $f(w)$, and $g(w)$ are given by (11)–(13).

Setting $\tau = 2$ leads to the graphs on the right of Fig. 2 and the results discussed at the beginning of this section.

2.4. *Likelihood of a phylogenetic tree*

A phylogenetic tree is a directed acyclic graph (DAG), which allows the expansion of the joint probability of the nodes in terms of the transition probabilities of (15). This expansion is based on the factorization rule for directed graphical models (see, for instance [12]), according to which the joint probability of a set of random variables $x_1, \ldots, x_n$ can be factorized as

$$P(x_1, \ldots, x_N) = \prod_{i=1}^{N} P(x_i|\text{parents}[x_i]), \tag{16}$$

where parents$[x_i]$ is the set of random variables corresponding to the subset of nodes with an arrow that feeds into x_i.

Consider Fig. 4, left. The black nodes, labelled by y_1, y_2, y_3, and y_4, represent contemporary species. The white nodes, labelled by z_1 and z_2, represent hypothetical ancestors. We are interested in the probability $P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S)$, where y_1, y_2, etc. represent nucleotides at the

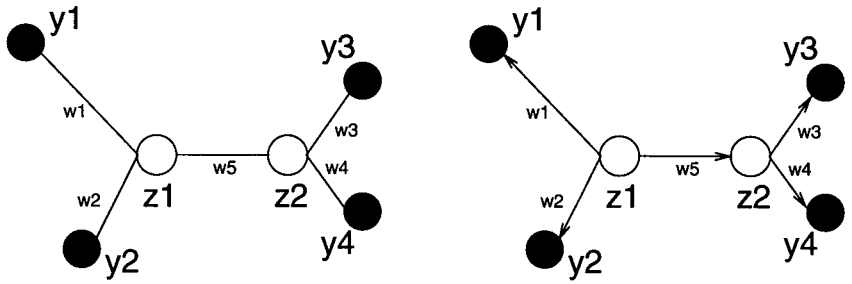

Fig. 4. **Phylogenetic trees.** Black nodes represent contemporary or extant species. White nodes represent hypothetical ancestors, where lineages bifurcate (speciation). *Left:* Undirected graph. *Right:* Directed graph. Node z_1 is the root of the tree, and arrows are directed.

respective nodes, $\mathbf{w}$ is the vector of all branch lengths, and S is a label defining the tree topology. Choosing, arbitrarily, z_1 to be the root of the tree, see Fig. 4, right, the application of (16) gives:

$$P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S) = P(y_1|z_1, w_1)P(y_2|z_1, w_2)P(z_2|z_1, w_5)P(y_3|z_2, w_3)P(y_4|z_2, w_4)\Pi(z_1). \tag{17}$$

The equilibrium distribution over the four nucleotides, $\Pi(z_1)$, is a parameter vector of the model. For example, in the Kimura model we have $\Pi(z_1 = A) = \Pi(z_1 = C) = \Pi(z_1 = G) = \Pi(z_1 = T) = 0.25$. The other factors represent transition probabilities, which are defined in (15).

Now, we assume that the transition matrix (15) is reversible:

$$P(Y|X, w)\Pi(X) = P(X|Y, w)\Pi(Y), \tag{18}$$

where $X, Y \in \{A, C, G, T\}$. Obviously, this holds true for the Kimura model discussed above. It can then be shown that the expansion of the joint probability $P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S)$ is independent of the root position.

Compare, for instance, the three directed graphs in Fig. 5. We have just derived the expansion for the tree on the left; see (17). Applying the expansion rule (16) to the tree in the middle, we obtain:

$$P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S) = P(y_1|z_1, w_1)P(y_2|z_1, w_2)P(z_1|z_2, w_5)P(y_3|z_2, w_3)P(y_4|z_2, w_4)\Pi(z_2). \tag{19}$$

Now, reversibility implies that $P(z_1|z_2, w_5)\Pi(z_2) = P(z_2|z_1, w_5)\Pi(z_1)$, hence the expansions in (17) and (19) are identical. By the same token,

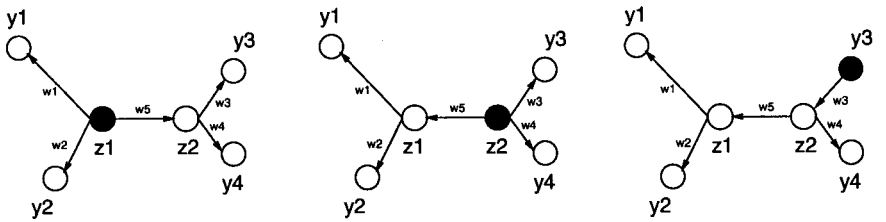

Fig. 5. **Different root positions.** The figure shows three directed graphs with different root positions (shown in black).

expanding the joint probability $P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S)$ according to the tree on the right of Fig. 5 gives

$$P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S) = P(y_1|z_1, w_1)P(y_2|z_1, w_2)P(z_1|z_2, w_5)P(y_4|z_2, w_4)P(z_2|y_3, w_3)\Pi(y_3). \tag{20}$$

Applying reversibility, $P(z_2|y_3, w_3)\Pi(y_3) = P(y_3|z_2, w_3)\Pi(z_2)$, this expansion is seen to be identical to (19) and hence (17). In the terminology of graphical models, the three directed graphs in Fig. 5 are *distribution equivalent* [12], that is, they represent the same joint probability distribution. In fact, a more rigorous proof [6] generalizes this finding to any phylogenetic tree: if the transition matrix is reversible, trees that only differ with respect to the position of the root and the directions of the arcs are equivalent. Consequently, we can choose the position of the root arbitrarily.[b]

The factorization (17) allows us to compute the probability of a complete configuration of nucleotides. However, while we obtain the nucleotides of the extant species, y_i, from the DNA sequence alignment, the nucleotides at the inner nodes, z_i, are never observed. This requires us to marginalize over them, as illustrated in Fig. 6:

$$P(y_1, y_2, y_3, y_4|\mathbf{w}, S) = \sum_{z_1}\sum_{z_2} P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S). \tag{21}$$

There are efficient message-passing algorithms to carry out this marginalization and decrease the computational complexity of the summation; see [6].

The upshot of this procedure is that for a given column $\mathbf{y}_t$ in the alignment, a probability $P(\mathbf{y}_t|\mathbf{w}, S)$ can be computed, which depends on the

[b]In more recent phylogenetic models, this reversibility constraint has been relaxed. See, for instance [8].

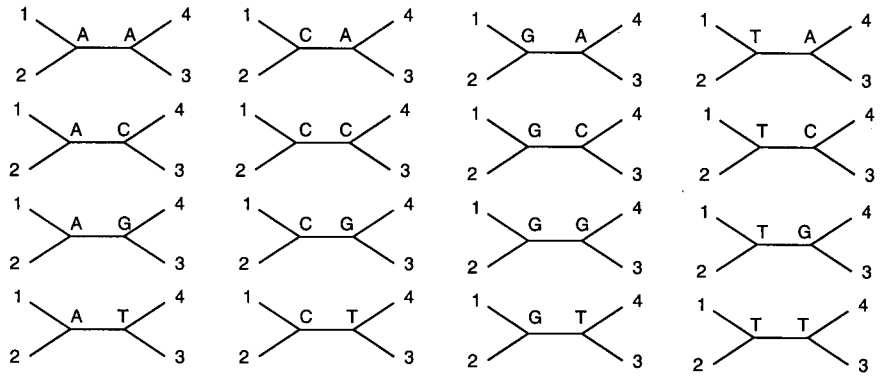

Fig. 6. **Marginalization over hidden nodes.** Leaf nodes represent extant taxa, which are observed (nucleotides in the DNA sequence alignment). Hidden nodes represent hypothetical ancestors, which are *not* observed (nuisance parameters). To obtain the probability of an observation, that is, the probability of observing a given column of nucleotides at a certain position in the DNA sequence alignment, one has to sum over all possible configurations of hidden nodes.

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, under the assumption that mutation events at different sites are independent of each other, the computation of the likelihood $P(\mathcal{D}|\mathbf{w}, S)$ of the whole DNA sequence alignment $\mathcal{D}$:

$$P(\mathcal{D}|\mathbf{w}, S) = \prod_{t=1}^{N} P(\mathbf{y}_t|\mathbf{w}, S). \tag{22}$$

This, in principle, opens the way to a maximum likelihood optimization of the tree: given a DNA sequence alignment $\mathcal{D}$, the tree $(\hat{S}, \hat{\mathbf{w}})$ most supported by the data is the one that maximizes the likelihood:

$$(\hat{S}, \hat{\mathbf{w}}) = \underset{S,\mathbf{w}}{\operatorname{argmax}}\{P(\mathcal{D}|\mathbf{w}, S)\}. \tag{23}$$

More precisely, one should also state the dependence of the likelihood on the nucleotide substitution model and its parameters, which also need to be optimized so as to maximize the likelihood. For the Kimura model, discussed above, we have one parameter: the transition-transversion ratio τ. Two more complex model, the HKY85 model [11] and the Felsenstein 84 model [5], have three further parameters: the equilibrium probabilities for the nucleotides, $\Pi(A), \Pi(C), \Pi(G), \Pi(T)$ (due to the constraint $\Pi(A) + \Pi(C) + \Pi(G) + \Pi(T) = 1$, there are three rather than four free parameters). Recently, more complex nucleotide substitution models have been

developed, as reviewed in [23]. These details are beyond the scope of this article. To keep the notation simple, the dependence of the likelihood on the nucleotide substitution model will not be stated explicitly.

A principled difficulty in applying the maximum likelihood method outline here is that the optimization problem is NP hard. As mentioned in Sec. 2.1, n taxa give rise to $(2n-5)!!$ different (unrooted) tree topologies, that is, the number of different tree topologies increases super-exponentially with the number of taxa. In practice this means that for large numbers of taxa one has to resort to iterative, greedy search algorithms, which usually find only a local rather than the global maximum of the likelihood. Effective algorithms have been proposed in [6] and [5], and are implemented in the program DNAML of the PHYLIP software package [7]. For an introductory text, see also [4]. The details of these optimization algorithms will not be summarized here. Instead, this article will focus on a fundamental problem inherent in the phylogenetic analysis of certain bacteria and viruses.

3. Recombination

Conventional phylogenetic analysis, as described in the previous section, assumes that all sites in a DNA multiple alignment have the same evolutionary history. This is a reasonable approach when applied to DNA sequences obtained from most species. However, this assumption is violated in certain bacteria and viruses due to interspecific *recombination*, which is a process that leads to the transfer or exchange of DNA subsequences between different strains. The resulting mixing of the genetic material by the formation of so-called *mosaic* sequences is likely to be an important source of genetic variation and is a process through which, for example, disease-causing bacteria may acquire resistance to antibiotics. Figure 8 shows an example in which the incorporation of the genetic material from another strain leads 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 mosaic sequences can lead to errors in phylogenetic tree estimation. Their detection, therefore, is a crucial prerequisite for inferring the evolutionary history of a set of DNA sequences.

Figure 9 shows an example of recombination in HIV-1 [25]. The left subfigure shows a phylogenetic tree for eight established strains of HIV-1. The subfigure on the right shows a so-called circulating recombinant strain, denoted by ZR-VI 191. If the phylogenetic analysis is done on the basis of the *env* gene, this strain is found to be most closely related to the A strain.

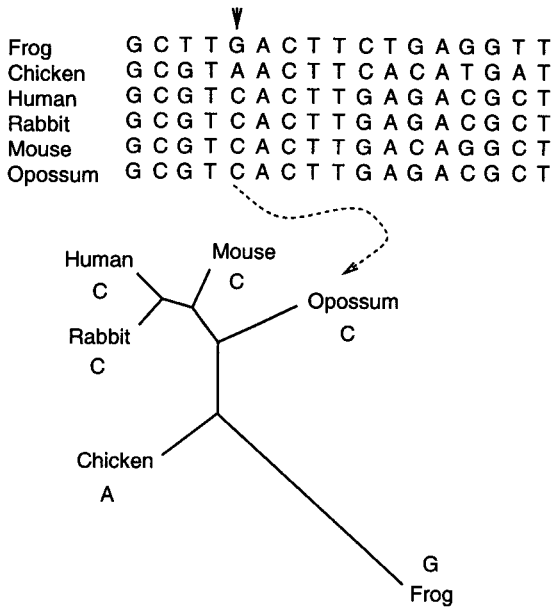

Fig. 7. **Statistical approach to phylogenetics.** For a given column $\mathbf{y}_t$ in the alignment, a probability $P(\mathbf{y}_t|\mathbf{w}, S)$ can be computed, which depends on the tree topology, S, and the vector of branch lengths, $\mathbf{w}$. This can be done for every site, $1 \leq t \leq N$, which allows the computation of the likelihood $P(\mathcal{D}|\mathbf{w}, S)$ of the whole DNA sequence alignment $\mathcal{D} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$.

For a phylogenetic analysis based on the *gag* gene, ZR-VI 191 is most closely related to the G strain. Ignoring recombination and treating the sequence of ZR-VI 191 as a monolithic entity will adversely affect the estimation of the branch lengths in the phylogenetic tree. For medical applications, determining a strain as a mosaic sequence of well-established strains can be important for vaccine development [25].

In the last few years, a plethora of methods for detecting interspecies recombination have been developed — following up on the seminal paper by John Maynard Smith [16] — and it is beyond the scope of this article to present a comprehensive overview. Many detection methods for identifying the nature and the breakpoints of the resulting mosaic structure are based on moving a window along the alignment and computing a phylogenetic divergence score for each window position. Examples are the bootstrap support for the locally optimal topology [26], the likelihood ratio between the locally and globally optimal trees [9], and the difference in the fitting

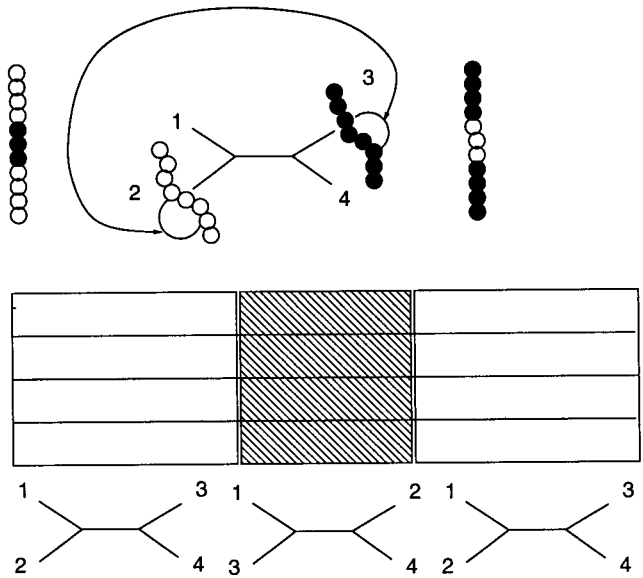

Fig. 8. **Influence of recombination on phylogenetic inference.** The figure shows a hypothetical phylogenetic tree of four strains. Recombination is the exchange of DNA subsequences between different strains (top diagram, middle), which results in two so-called mosaic sequences (top diagram, margins). The affected region in the multiple DNA sequence alignment (shown by the shaded area in the middle diagram) seems to originate from a different phylogenetic topology, in which two branches of the phylogenetic tree have been exchanged (bottom diagram, where the numbers at the leaves represent the four strains). Reprinted from [14], with permission from Mary Ann Liebert.

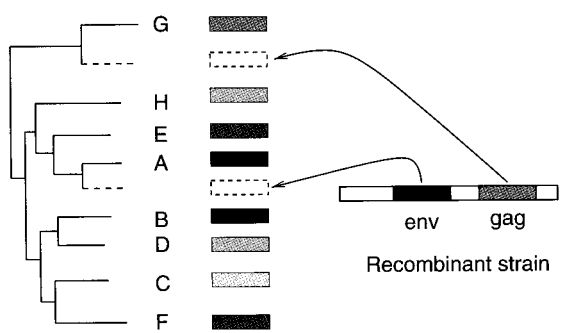

Fig. 9. **Recombination in HIV-1.** The left subfigure shows a phylogenetic tree for eight established strains of HIV-1. The subfigure on the right shows a so-called circulating recombinant strain, denoted by ZR-VI 191. If the phylogenetic analysis is done on the basis of the *env* gene, this strain is found to be most closely related to the A strain. For a phylogenetic analysis based on the *gag* gene, ZR-VI 191 is most closely related to the G strain.

scores between two adjacent locally optimized trees [17]. The determination of the breakpoints of the mosaic structure is then based on an analysis of the signals thus obtained, using bootstrapping to estimate their significance. While these methods are useful for a preliminary scan of a DNA sequence alignment, the spatial resolution for the identification of the breakpoints is typically of the order of the window size and, consequently, rather poor.

This chapter discusses a different approach, which was first suggested in [13]. The idea is to introduce a hidden state, which 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, a site graph is introduced, representing which nucleotides interact in determining the tree topology. To keep the mathematical model tractable and the computational costs limited, interactions are reduced to nearest-neighbour interactions. The natural framework for modelling such a system is a hidden Markov model, whose application to the detection of recombination was first suggested in [18]. The next section provides a brief introduction to hidden Markov models.

4. A One-Minute Introduction to Hidden Markov Models

Assume you are in a casino and take part in some (hopefully legal) gambling game involving a die. You are playing against two players: a fair player, who uses a fair die, and a corrupt player, who uses a loaded die. The situation is illustrated in Fig. 10. Unfortunately, the other players are hidden behind a brick wall, and all you observe is a sequence of die faces; see Fig. 11. The

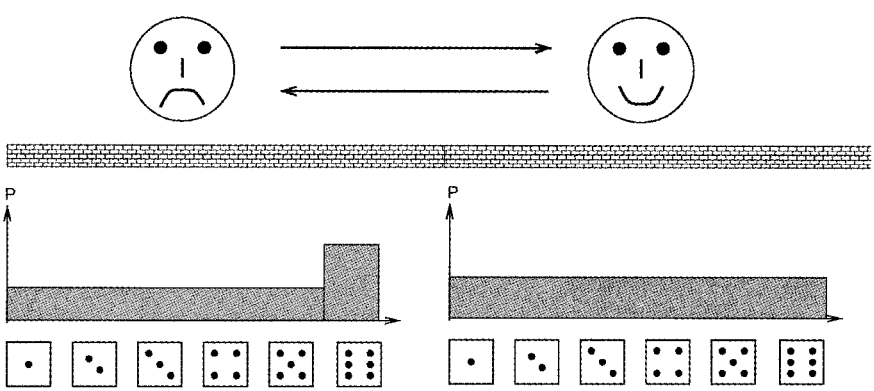

Fig. 10. **Corrupt casino 1.** Two players are in a casino: a fair player (right) using a fair die, and a corrupt player (left) using a loaded die.

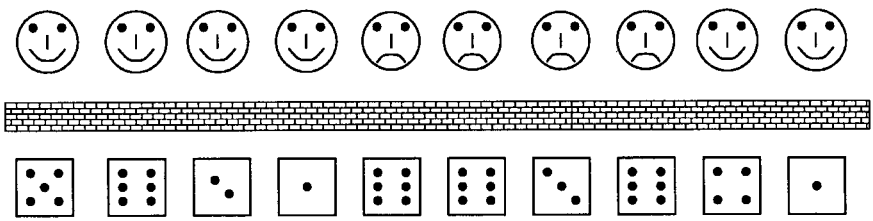

Fig. 11. **Corrupt casino 2.** The player is hidden behind a brick wall, and only the die faces are observed. The problem is to predict which player is rolling the die at a given time t.

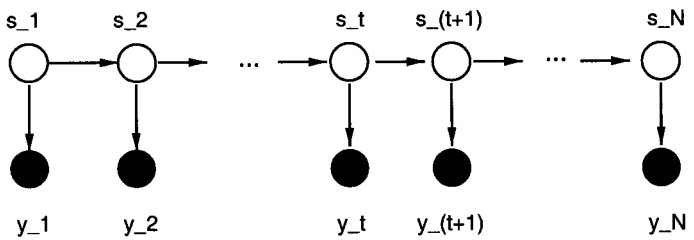

Fig. 12. **Hidden Markov model.** Black nodes represent observed random variables (the die faces), white nodes represent hidden states (the players), and arcs represent conditional dependencies. The joint probability factorizes into a product of emission probabilities (vertical arrows) and transition probabilities (horizontal arrows). The prediction task is to find the most likely sequence of hidden states given the observations.

task is to predict which player is rolling the die at a given time, and to predict the breakpoint where the corrupt player is taking over (in order to nab him).

If the decision of a player to pass the die on to the other player is made instantaneously on the basis of the current situation without considering the earlier past, the process corresponds to a hidden Markov model (HMM), shown in Fig. 12. Here, black nodes represent observed random variables y_t (the die faces) at different moments in time t, white nodes represent hidden states S_t (the players) at different times, and arcs represent conditional dependencies. The task is to find the most likely sequence of hidden states given the observations, that is, the mode of

$$P(\mathbf{S}|\mathbf{y}) = P(S_1, \ldots, S_N | y_1, \ldots, y_N). \tag{24}$$

At first, this task seems to be intractable: for K different states (here: $K = 2$ for "fair" and "corrupt") and a sequence of length N, there are

K^N different state sequences. Hence, an exhaustive search seems to be impossible for all but very short sequence lengths N. Fortunately, there is a dynamic programming method, the so-called *Viterbi algorithm*, which reduces the computational complexity to $\mathcal{O}(N)$ (that is, linear in N) by exploiting the sparseness of the connectivity of the graph in Fig. 12.

Recall that in a directed graphical model the joint probability of the random variables $x_1, \ldots, x_N$ can be factorized according to (16). The application of this formula to the graph in Fig. 12 gives:

$$P(y_1, \ldots, y_N, S_1, \ldots, S_N) = \prod_{t=1}^{N} P(y_t|S_t) \prod_{t=2}^{N} P(S_t|S_{t-1}) P(S_1). \tag{25}$$

We refer to $P(y_t|S_t)$ as the *emission probabilities* (corresponding to the vertical edges), $P(S_t|S_{t-1})$ as the *transition probabilities* (which correspond to the horizontal edges), and $P(S_1)$ as the *initial probability*. From (25) we obtain the recursion:

$$\begin{aligned}
\gamma_n(S_n) &= \max_{S_1,\ldots,S_{n-1}} \ln P(y_1, \ldots, y_n, S_1, \ldots, S_n) \\
&= \max_{S_1,\ldots,S_{n-1}} \left[\sum_{t=1}^{n} \ln P(y_t|S_t) + \sum_{t=2}^{n} \ln P(S_t|S_{t-1}) + \ln P(S_1) \right] \\
&= \ln P(y_n|S_n) + \max_{S_{n-1}} \left[\ln P(S_n|S_{n-1}) + \max_{S_1,\ldots,S_{n-2}} \left[\sum_{t=1}^{n-1} \ln P(y_t|S_t) \right.\right. \\
&\quad \left.\left. + \sum_{t=2}^{n-1} \ln P(S_t|S_{t-1}) + \ln P(S_1) \right] \right] \\
&= \ln P(y_n|S_n) + \max_{S_{n-1}} [\ln P(S_n|S_{n-1}) + \gamma_{n-1}(S_{n-1})].
\end{aligned} \tag{26}$$

Obviously:

$$\begin{aligned}
\max_{S_1,\ldots,S_N} P(S_1, \ldots, S_N|y_1, \ldots, y_N) &= \max_{S_1,\ldots,S_N} \ln P(y_1, \ldots, y_N, S_1, \ldots, S_N) \\
&= \max_{S_N} \gamma_N(S_N)
\end{aligned} \tag{27}$$

and the mode, $P(\hat{S}_1, \ldots, \hat{S}_N|y_1, \ldots, y_N)$, is obtained by recursive backtracking:

Initialization:

$$\hat{S}_N = \operatorname*{argmax}_{S_N} \gamma_N(S_N). \tag{28}$$

Recursion:

$$\hat{S}_{n-1} = \underset{S_{n-1}}{\operatorname{argmax}} \left[\ln P(\hat{S}_n | S_{n-1}) + \gamma_{n-1}(S_{n-1}) \right]. \tag{29}$$

The computational complexity of a single step of the recursions (26) and (29) is $\mathcal{O}(K^2)$, that is, it only depends on the number of different states K but is independent of the sequence length N. The total computational complexity of the algorithm is thus linear in N, which is a considerable improvement over the naive method, which was K^N. For a more detailed exposition of this topic, see [24].

5. Detecting Recombination with Hidden Markov Models

5.1. *The model*

Let us now study how HMMs can be applied to model mosaic structures in DNA sequence alignments. Here, the hidden state represents the phylogenetic tree topology at a given site. For four taxa, for instance, there are three possible tree topologies, shown in Fig. 13. The subscript t now represents sites in the DNA sequence alignment rather than time, hence S_t is the hidden state corresponding to the tth site in the alignment. The observations $\mathbf{y}_t$ are the columns of the DNA sequence alignment, that is, $\mathbf{y}_t$ is the vector with the nucleotides of all the taxa at the tth site in the alignment. For a given tree, we can compute the probability of $\mathbf{y}_t$, as discussed in Sec. 2.4 and illustrated in Fig. 7. Hence for a given DNA sequence alignment $\mathcal{D} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, we can apply the Viterbi algorithm to find the most likely sequence of hidden states, $S_1, \ldots, S_N$, that is, the mode of $P(S_1, \ldots, S_N | \mathbf{y}_1, \ldots, \mathbf{y}_N)$. Recombination events then correspond to state transitions in the Viterbi path.

Recall that in an HMM, the joint probability factorizes into the product of the emission probabilities, $P(\mathbf{y}_t | S_t)$, and the transition probabilities, $P(S_t | S_{t-1})$, where the latter correspond to recombination events. With K

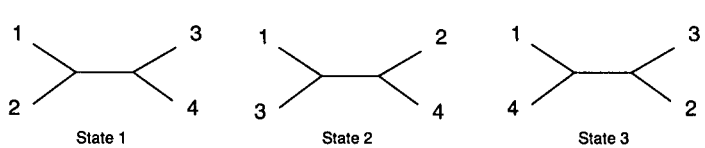

Fig. 13. **Different tree topologies for four taxa.** Shown are the three possible phylogenetic tree topologies for four taxa. Species 1 can be clustered with species 2, 3, or 4. Reprinted from [14], with permission from Mary Ann Liebert.

different tree topologies, there are, in principle, $K(K-1)$ transition probabilities to be specified. However, given that recombination is likely to be a rare event, it would hardly be possible to reasonably infer these parameters from the DNA sequence alignment (over-fitting), nor is it likely that detailed prior knowledge is available to decide on these parameters in advance. For this reason, only *one* free parameter was used in [18]: the overall probability that no recombination occurs. This is similar to an approach taken in [5] for modelling rate variation among sites. Let ν be the probability that the tree topology remains unchanged as we move from a given site in the alignment, t, to an adjacent site, $t+1$ or $t-1$. We then obtain for the state transition probabilities:

$$\begin{aligned} P(S_t|S_{t-1}) &= \nu\delta(S_t, S_{t-1}) + \frac{1-\nu}{K-1}[1-\delta(S_t, S_{t-1})] \\ &= \nu^{\delta(S_t,S_{t-1})}\left(\frac{1-\nu}{K-1}\right)^{1-\delta(S_t,S_{t-1})}, \end{aligned} \tag{30}$$

where $\delta(S_t, S_{t-1})$ denotes the Kronecker delta function, which is 1 when $S_t = S_{t-1}$, and 0 otherwise. It is easily checked that this satisfies the normalization constraint $\sum_{S_t} P(S_t|S_{t-1}) = 1$. For the emission probabilities, recall from Sec. 2.4 and Fig. 7 that for a given nucleotide substitution model, the probability of a column vector $\mathbf{y}_t$ depends both on the tree topology, S_t, and the vector of branch lengths corresponding to this topology, $\mathbf{w}_{S_t}$. To simplify the notation, let us introduce the accumulated vector of all branch lengths in all possible topologies, $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_K)$, and define: $P(\mathbf{y}_t|S_t, \mathbf{w}_{S_t}) = P(\mathbf{y}_t|S_t, \mathbf{w})$. This means that S_t indicates which subvector of $\mathbf{w}$ applies. We can depict the dependence of the probability distribution on the parameters $\mathbf{w}$ and ν in an extended graphical model, shown in Fig. 14. Applying the Viterbi algorithm gives us the most likely hidden state sequence conditional on the observations (that is, the DNA sequence alignment) and the parameters $\mathbf{w}$ and ν:

$$\underset{S_1,\ldots,S_N}{\operatorname{argmax}} P(S_1, \ldots, S_N|\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{w}, \nu). \tag{31}$$

We thus need a way to estimate these parameters.

5.2. *Naive parameter estimation*

A straightforward way to estimate the branch lengths $\mathbf{w}$ seems to be a separate maximum likelihood optimization for each possible tree topology. This can be accomplished with the methods described at the end of Sec. 2.4, and was applied in [18]. However, Fig. 8 points to a serious shortcoming

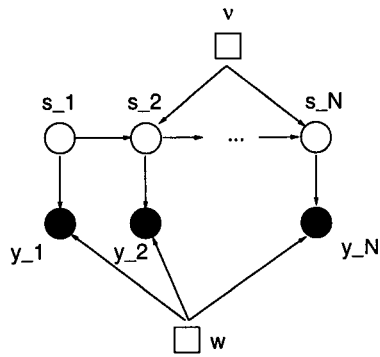

Fig. 14. **Modelling recombination with a hidden Markov model.** Positions in the model, labelled by the subscript t, correspond to positions in the DNA sequence alignment. Black nodes represent observed random variables; these are the columns in the DNA sequence alignment. White nodes represent hidden states; these are the different tree topologies, as shown in Fig. 13. Squares represent parameters of the model: the vector of branch lengths $\mathbf{w}$, and the recombination parameter ν. Arcs represent conditional dependencies. 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 and the vector of branch lengths $\mathbf{w}$. 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.

of this approach. For a proper estimation of the branch lengths of the recombinant tree, that is, the tree that corresponds to the shaded centre region of the alignment, one would have to base the parameter estimation on this very region of the alignment. Unfortunately, its location is not known in advance. Estimating the branch lengths from the whole DNA sequence alignment leads to seriously distorted values — see Fig. 15 — since the estimation includes data for which the tree topology is incorrect. A heuristic way to address this problem, suggested in [18], is to estimate the branch lengths from a subregion of the alignment. The length of this region should be matched to the length of the recombinant region, which, however, is not known in advance. Also, this approach does not offer a way to estimate the recombination parameter ν.

5.3. *Maximum likelihood*

A solution to this problem, proposed in [14], is a proper maximum likelihood estimation of the parameters so as to maximize

$$L(\mathbf{w}, \nu) = \ln P(\mathcal{D}|\mathbf{w}, \nu) = \ln \sum_{\mathbf{S}} P(\mathcal{D}, \mathbf{S}|\mathbf{w}, \nu) \tag{32}$$

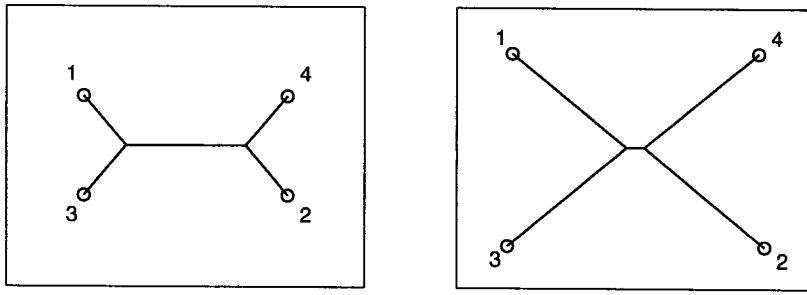

Fig. 15. **Effect of naive parameter estimation.** The left figure shows the correct recombinant tree, corresponding to the recombinant region in the alignment of Fig. 8. The right figure shows the tree that results from a maximum likelihood estimation of the branch lengths from the whole DNA sequence alignment. This includes the flanking regions — shown in white in Fig. 8 — where the recombinant tree topology is incorrect. Obviously, the branch lengths have been significantly distorted, with a contraction of the internal branch and an extension of the external branches. Reprinted from [14], with permission from Mary Ann Liebert.

with respect to the vector of branch lengths $\mathbf{w}$ and the recombination parameter ν. This requires a summation over all state sequences $\mathbf{S} = (S_1, \ldots, S_N)$, that is, over K^N terms. For all but very short sequence lengths N this is intractable. A viable alternative, however, is the expectation maximization (EM) algorithm [3]. Let $Q(\mathbf{S})$ denote an arbitrary probability distribution over the hidden state sequences, and define

$$U(\mathbf{w}, \nu) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln P(\mathcal{D}, \mathbf{S}|\mathbf{w}, \nu) - \sum_{\mathbf{S}} Q(\mathbf{S}) \ln Q(\mathbf{S}). \tag{33}$$

We are interested in the posterior distribution of the hidden state sequences, $P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \nu)$, given the DNA sequence alignment, $\mathcal{D}$, and the parameters, $\mathbf{w}$ and ν. The difference between $Q(\mathbf{S})$ and $P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \nu)$ is measured by the Kullback–Leibler divergence

$$KL(Q, P) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \left(\frac{Q(\mathbf{S})}{P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \nu)} \right), \tag{34}$$

which is always non-negative, and zero if and only if $Q = P$. The proof, which is based on the concavity of the logarithm, is straightforward. Now, combining (33) and (34), we can rewrite the likelihood of (32) as

$$L(\mathbf{w}, \nu) = U(\mathbf{w}, \nu) + KL(Q, P). \tag{35}$$

This decomposition was first suggested in [20], and can easily be proved by recalling that $P(\mathcal{D}, \mathbf{S}|\mathbf{w}, \nu) = P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \nu) P(\mathcal{D}|\mathbf{w}, \nu)$ and $\sum_{\mathbf{S}} Q(\mathbf{S}) = 1$.

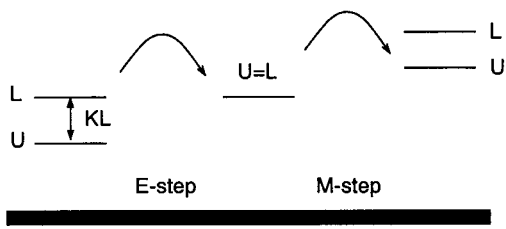

Fig. 16. **Illustration of the EM algorithm.** U is a lower bound on the log likelihood L, with a difference given by the Kullback–Leibler divergence KL. The E-step sets KL to zero. Since the model parameters are kept constant, the log likelihood L is not changed. The M-step adapts the model parameters so as to maximize U. Since U is a lower bound on L, this also increases L.

Since $KL(Q,P)$ is non-negative, U is a lower bound on L: $U(\mathbf{w},\nu) \leq L(\mathbf{w},\nu)$. The EM algorithm alternates between optimizing the distribution over the hidden states $Q(\mathbf{S})$ (the E-step) and optimizing the parameters given $Q(\mathbf{S})$ (the M-step). The E-step holds the parameters fixed and sets Q to the posterior distribution over the hidden states given the parameters, $Q(\mathbf{S}) = P(\mathbf{S}|\mathcal{D},\mathbf{w},\nu)$. This sets $KL(Q,P) = 0$ and, consequently, $L(\mathbf{w},\nu) = U(\mathbf{w},\nu)$. The M-step holds the distribution $Q(\mathbf{S})$ fixed and computes the parameters $\mathbf{w},\nu$ that maximize U. Since $L(\mathbf{w},\nu) = U(\mathbf{w},\nu)$ at the beginning of the M-step, and since the E-step does not affect the model parameters, each EM cycle is guaranteed to increase the likelihood unless the system has already converged to a (local) maximum (or, less likely, a saddle point). An illustration of the algorithm is given in Fig. 16.

Now, similar to the discussion in Sec. 4, we can exploit the sparseness of the connectivity of the underlying graphical model and simplify the maximization of U considerably. From the factorization (25) we have:

$$\begin{aligned} P(\mathcal{D},\mathbf{S}|\mathbf{w},\nu) &= P(\mathbf{y}_1,\ldots,\mathbf{y}_N,S_1,\ldots,S_N|\mathbf{w},\nu) \\ &= \prod_{t=1}^{N} P(\mathbf{y}_t|S_t,\mathbf{w}) \prod_{t=2}^{N} P(S_t|S_{t-1},\nu) P(S_1). \end{aligned} \tag{36}$$

Inserting (36) into (33) gives

$$\begin{aligned} U(\mathbf{w},\nu) = &\sum_{\mathbf{S}} Q(\mathbf{S}) \sum_{t=1}^{N} \ln P(\mathbf{y}_t|S_t,\mathbf{w}) \\ &+ \sum_{\mathbf{S}} Q(\mathbf{S}) \sum_{t=2}^{N} \ln P(S_t|S_{t-1},\nu) + C, \end{aligned} \tag{37}$$

where C is independent of the parameters $\mathbf{w}$ and ν. Equation (37) simplifies considerably. The first term allows immediate marginalization over all but one state S_t in the state sequence $\mathbf{S}$:

$$\sum_{S} Q(\mathbf{S}) \sum_{t=1}^{N} \ln P(\mathbf{y}_t|S_t, \mathbf{w}) = \sum_{t=1}^{N} \sum_{S_t=1}^{K} Q(S_t) \ln P(\mathbf{y}_t|S_t, \mathbf{w}). \tag{38}$$

For the second term, recall the definition of the transition probabilities $P(S_t|S_{t-1}, \nu)$ in (30), define

$$\Psi = \sum_{\mathbf{S}} \sum_{t=2}^{N} Q(\mathbf{S}) \delta(S_t, S_{t-1}) = \sum_{t=2}^{N} \sum_{S_t=1}^{K} Q(S_t, S_{t-1} = S_t) \tag{39}$$

and note that

$$\sum_{\mathbf{S}} \sum_{t=2}^{N} Q(\mathbf{S})[1 - \delta(S_t, S_{t-1})] = N - 1 - \Psi. \tag{40}$$

This gives

$$\sum_{\mathbf{S}} Q(\mathbf{S}) \sum_{t=2}^{N} \ln P(S_t|S_{t-1}, \nu) = \Psi \ln \nu + (N - 1 - \Psi) \ln \left(\frac{1 - \nu}{K - 1} \right). \tag{41}$$

Inserting (38) and (41) into (37), we obtain:

$$\begin{aligned} U = & \sum_{t=1}^{N} \sum_{S_t=1}^{K} Q(S_t) \ln P(\mathbf{y}_t|S_t, \mathbf{w}) + \Psi \ln \nu \\ & + (N - 1 - \Psi) \ln \left(\frac{1 - \nu}{K - 1} \right) + C. \end{aligned} \tag{42}$$

Note that U only depends on the marginal univariate probability $Q(S_t)$, and the marginal two-variate probability $Q(S_t, S_{t-1})$ (via (39)), but no longer on the multivariate joint probability $Q(\mathbf{S})$.

5.3.1. *E-step*

The probabilities $Q(S_t)$ and $Q(S_t, S_{t+1})$ are updated in the E-step, where we set:

$$Q(S_t) \rightarrow P(S_t|\mathcal{D}, \mathbf{w}, \nu) \tag{43}$$

$$Q(S_{t-1}, S_t) \rightarrow P(S_{t-1}, S_t|\mathcal{D}, \mathbf{w}, \nu). \tag{44}$$

These computations are carried out with the *forward-backward* algorithm for HMMs [24], which is a dynamic programming method that reduces the

computational complexity from $O(K^N)$ to $O(N)$. The underlying principle is similar to that of the Viterbi algorithm, discussed in Sec. 4, and is based on the sparseness of the connectivity in the HMM structure. Details are beyond the scope of this chapter, and the interested reader is referred to the tutorial [24], or textbooks like [1] and [4], which also discuss implementation issues.

Now, all that remains to be done is to derive update equations for the parameters $\mathbf{w}$ and ν so as to maximize the function U (M-step).

5.3.2. *M-step: Optimization of the recombination parameter*

Setting the derivative of U with respect to ν to zero, $\frac{\partial U}{\partial \nu} = 0$, we obtain

$$\nu = \frac{\Psi}{N-1}. \tag{45}$$

This optimization is straightforward since, as seen from (39), Ψ only depends on $Q(S_{t-1}, S_t)$, which is obtained by application of the forward-backward algorithm (see above).

5.3.3. *M-step: Optimization of the branch lengths*

Only the first term on the left-hand side of (42) depends on the branch lengths $\mathbf{w}$. This requires a maximization of

$$\sum_{t=1}^{N} \sum_{S_t=1}^{K} Q(S_t) \ln P(\mathbf{y}_t | S_t, \mathbf{w}), \tag{46}$$

which can be achieved with standard phylogenetic programs, like PHYLIP (mentioned in Sec. 2.4). The only modification required is the introduction of a weighting factor $Q(S_t)$ for each site, as illustrated in Fig. 17.

5.3.4. *Reason for not optimizing the prior probabilities*

In principle, U has a further set of parameters that need to be optimized: the $K-1$ prior probabilities $P(S_1)$ (see (36)). Due to the rarity of recombination events, however, a maximum likelihood approach would most probably lead to over-fitting. Also, since DNA sequence alignments are usually sufficiently long, $N \gg K$, the influence of $P(S_1)$ on the mode of $P(S_1, \ldots, S_N | \mathcal{D})$ is negligible. It therefore seems to be reasonable to keep the prior probabilities constant: $P(S_1) = \frac{1}{K} \forall S_1 \in \{1, \ldots, K\}$.

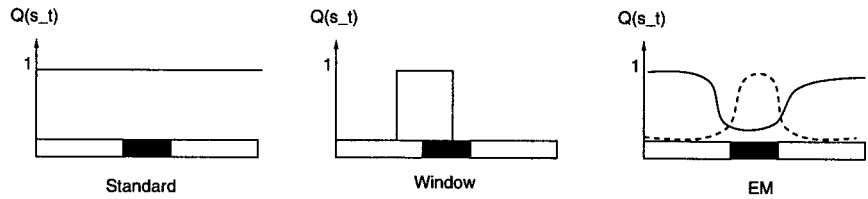

Fig. 17. **Nucleotide weighting schemes.** The figure shows three nucleotide weighting schemes for estimating the branch lengths of the phylogenetic trees. The bottom of each figure represents a multiple DNA sequence alignment with a recombinant zone, printed in grey, in the middle. *Left:* Naive approach, suggested in [18], where the tree parameters are estimated from the whole alignment. This corresponds to constant weights, $Q(S_t) = 1 \; \forall t$. *Middle:* Heuristic window method, also suggested in [18], where the tree parameters are estimated from a subregion of the alignment. The length of this region should be matched to the length of the recombinant region, which, however, is not known in advance. *Right:* Maximum likelihood with the EM algorithm. The dashed line shows the site-dependent weights $Q(S_t = T_R)$ for the recombinant topology T_R, the solid line represents the weights for the non-recombinant topology T_0 : $Q(S_t = T_0)$. Note that in this scheme the weights $Q(S_t)$ are updated automatically in every iteration of the algorithm as a natural consequence of the optimization procedure (E-step). Reprinted from [14], with permission from Mary Ann Liebert.

5.3.5. *Algorithm*

The implementation of the parameter update scheme is straightforward and can be accomplished with the following algorithm:

(1) Initialize the parameters $\mathbf{w}$ and ν. This can be done as in [18], that is, by choosing a plausible recombination rate and by estimating $\mathbf{w}$, for each of the topologies, with a phylogenetic program like DNAML from the whole alignment.
(2) Compute $Q(S_t)$ and $Q(S_{t-1}, S_t)$ with the forward-backward algorithm for HMMs.
(3) Compute Ψ from (39) and adapt ν according to (45).
(4) For $t = 1$ to N: weight the tth column in the multiple sequence alignment, $\mathbf{y}_t$, by $Q(S_t)$, and optimize the branch lengths $\mathbf{w}$ so as to maximize $U(\mathbf{w})$ in (46). This can, in principle, be achieved with a standard phylogeny program, like DNAML of the PHYLIP package [7]. The only change required is the introduction of a weighting scheme for the sites in the alignment.
(5) Test for convergence. If the algorithm has not yet converged, go back to step 2.

Note that this algorithm can be interpreted as a modified version of the Baum–Welch algorithm; see [24].

6. Test Data

The viability of the proposed HMM scheme was tested on the following three DNA sequence alignments.

6.1. *Synthetic data*

DNA sequences, 1000 nucleotides long, were evolved along a 4-species tree, using the Kimura model of nucleotide substitution, which was described in Sec. 2.3. The transition-transversion ratio was set to $\tau = 2$. Two recombination events were simulated by exchanging the indicated lineages, as shown in Fig. 18.

6.2. *Gene conversion in maize*

When looking at the distribution of genes within genomes, one finds that many genes, rather than existing as individual copies, are part of a larger family of related genes called a *multigene family.* A special form of recombination, which takes place in multigene families and contributes greatly to their evolution, is *gene conversion.* This process occurs when the DNA sequence of one gene is replaced (or "converted") by the DNA sequence from another; for further details, see, for instance [21], Chapter 3. Evidence for gene conversion between a pair of maize actin genes (involving Maz56 and Maz63; see below) has been reported in [19]. In the present study, the following four maize sequences were analyzed: Maz56 (GenBank/EMBL accession number U60514), Maz63 (U60513), Maz89 (U60508), and Maz95 (U60507). As discussed in Sec. 2.2, prior to any phylogenetic analysis the DNA sequences need to be aligned. This was done with the program Clustal-W [28], using the default parameter settings and discarding columns with gaps. The three hidden states of the HMM are defined as follows. State 1: ((Maz56,Maz63),(Maz89,Maz95));

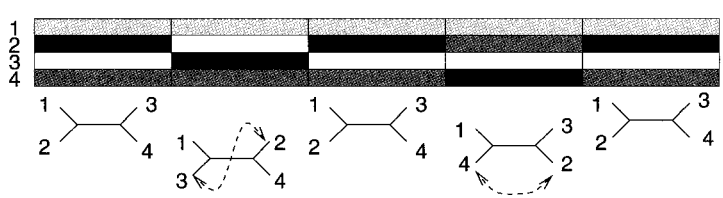

Fig. 18. **Synthetic DNA sequence alignment.** Two recombination events are simulated by swapping the indicated lineages. Defining the predominant tree topology as state 1, the first recombination event corresponds to a transition into state 2, while the second event corresponds to a transition into state 3.

state 2: ((Maz56,Maz89),(Maz63,Maz95)); state 3: ((Maz56,Maz95), (Maz63,Maz89)).

6.3. *Recombination in Neisseria*

One of the first indications for interspecific recombination was found in the bacterial genus *Neisseria* [16]. The analysis in this study was done on a subset of the 787 nucleotide *Neisseria argF* DNA multiple alignment studied in [29], selecting the following four strains: (1) *N. gonorrhoeae* (X64860), (2) *N. meningitidis* (X64866), (3) *N. cinera* (X64869), and (4) *N. mucosa* (X64873) (GenBank/EMBL accession numbers are in brackets). Zhou and Spratt [29] found two anomalous, or more diverged regions in the DNA alignment, which occur at positions $t = 1 - 202$ and $t = 507 - 538$.[c] In the rest of the alignment, *N. meningitidis* clusters with *N. gonorrhoeae* (defined as state 1 in our HMM), while between $t = 1$ and $t = 202$, they found that it is grouped with *N. cinera* (defined as state 3 in our HMM). Zhou and Spratt [29] suggested that the region $t = 507 - 538$ was more diverged as a result of rate variation. An illustration is given in Fig. 19.

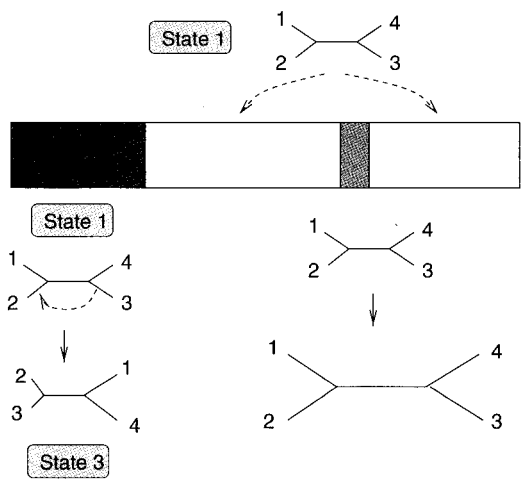

Fig. 19. **Recombination in Neisseria.** According to [29], a recombination event corresponding to a transition from state 1 into state 3 has affected the first 202 nucleotides of the DNA sequence alignment. A second more diverged region seems to be the result of rate variation.

[c]Note that Zhou and Spratt [29] used a different labeling scheme, with the first nucleotide at $t = 296$, and the last one at $t = 1082$.

7. Simulation

Both training schemes, the heuristic method described in Sec. 5.2, and the maximum likelihood approach described in Sec. 5.3, were tested on the three DNA sequence alignments. The application of the heuristic method was similar to [18]. For each of the three possible tree topologies, the branch lengths were estimated separately with maximum likelihood on the whole alignment, using the Kimura model of nucleotide substitution, which was described in Sec. 2.3. The practical computation was carried out with the program DNAML of the PHYLIP package [7]. The transition-transversion ratio τ was optimized with maximum likelihood, using the program package PUZZLE [27]. The recombination parameter was set to $\nu = 0.8$. As opposed to [18], the optimization was not restricted to subsets of the alignments, since the subset size is a parameter that cannot be properly optimized.

The maximum likelihood approach followed the procedure described in Sec. 5.3, optimizing all the parameters simultaneously with the EM algorithm. The initial recombination parameter was set to $\nu = 0.8$, as for the heuristic approach, and the initial probabilities for the three tree topologies were set to equal values: $P(S_1 = 1) = P(S_1 = 2) = P(S_1 = 3) = 1/3$. The EM algorithm typically took about 10–30 EM steps to converge, depending on the data set. Further details can be found in [14].

After parameter estimation, the classification of a site can be based on the mode of the posterior probability $P(S_t|\mathcal{D})$, that is, set $S_t = k$ if $P(S_t = k|\mathcal{D}) \geq P(S_t = i|\mathcal{D}) \forall i \neq k$. A problem of this approach is that even if $S_t = k_t$ maximizes $P(S_t|\mathcal{D})$ for all $t \in \{1, \ldots, N\}$, it is not guaranteed that $(k_1, k_2, \ldots, k_N)$ maximizes $P(S_1, S_2, \ldots, S_N|\mathcal{D})$ [24].[d] Therefore, a better approach is to base the classification of the sites S_t directly on the mode of the joint posterior probability $P(S_1, S_2, \ldots, S_N|\mathcal{D})$, which can be computed with the Viterbi algorithm, described in Sec. 4. However, the deviation between the predictions based on the mode of the marginal posterior probabilities $P(S_t|\mathcal{D})$ and the joint posterior probability $P(S_1, S_2, \ldots, S_N|\mathcal{D})$ was found to be negligible in the simulation studies described here, and the marginal posterior probability $P(S_t|\mathcal{D})$ has the advantage that it can be graphically displayed.

[d] Assume, for instance, that $S_t = k_t$ maximizes $P(S_t|\mathcal{D})$ and $S_{t+1} = k_{t+1}$ maximizes $P(S_{t+1}|\mathcal{D})$, but that $P(S_{t+1} = k_{t+1}|S_t = k_t) = 0$. Then $P(S_t = k_t, S_{t+1} = k_{t+1}|\mathcal{D}) = 0$, so (k_t, k_{t+1}) is not the mode of $P(S_t S_{t+1}|\mathcal{D})$.

This visualization has been done in Figs. 20–22, which show the results obtained with the two training methods on the three DNA sequence alignments. Each figure contains two subfigures: the left subfigure shows the results obtained with the heuristic training scheme, and the right subfigure shows the results obtained with the maximum likelihood scheme. Each subfigure is composed of three graphs. These graphs show the posterior probabilities for the three topologies, $P(S_t = 1|\mathcal{D})$ (top), $P(S_t = 2|\mathcal{D})$ (middle),

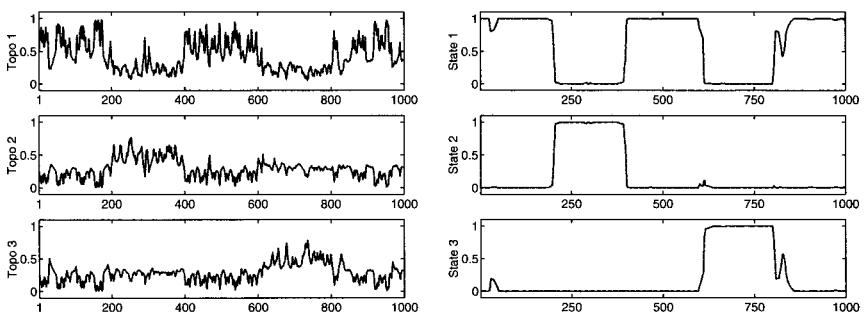

Fig. 20. **Detection of recombination in the synthetic DNA sequence alignment.** The figure contains two subfigures, where each subfigure is composed of three graphs. These graphs show the posterior probabilities for the three topologies, $P(S_t = 1|\mathcal{D})$ (top), $P(S_t = 2|\mathcal{D})$ (middle), $P(S_t = 3|\mathcal{D})$ (bottom), plotted along the DNA sequence alignment (the subscript t denotes the position in the alignment). *Left:* Heuristic training scheme. *Right:* Parameter estimation with maximum likelihood.

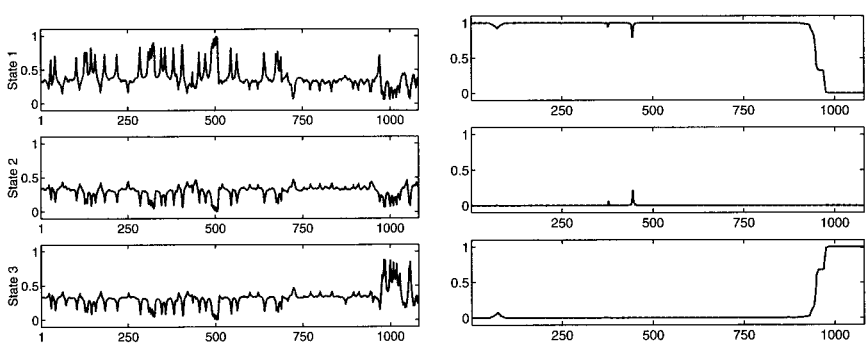

Fig. 21. **Detection of gene conversion between two maize actin genes.** The figure contains two subfigures, where each subfigure is composed of three graphs, as explained in the caption of Figure 20. *Left:* Heuristic training scheme. *Right:* Parameter estimation with maximum likelihood.

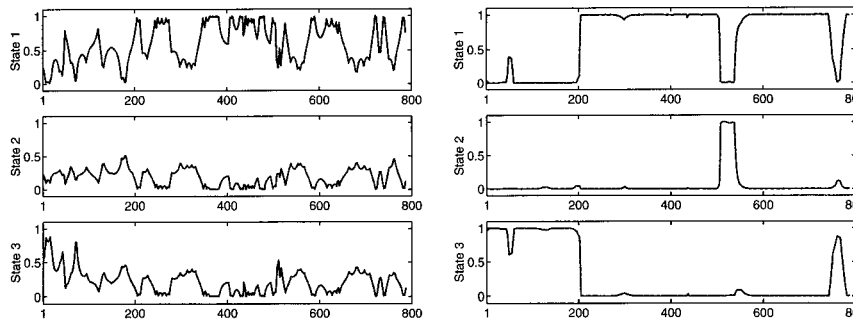

Fig. 22. **Detection of recombination in the Neisseria DNA sequence alignment.** The figure contains two subfigures, where each subfigure is composed of three graphs, as explained in the caption of Fig. 20. *Left:* Heuristic training scheme. *Right:* Parameter estimation with maximum likelihood.

and $P(S_t = 3|\mathcal{D})$ (bottom), plotted along the DNA sequence alignment (recall that the subscript t denotes the position in the alignment). The probabilities are computed with the forward-backward algorithm, which was mentioned in Sec. 5.3, and is discussed at length in [24].

7.1. *Synthetic DNA sequence alignment*

Figure 20 shows the results obtained on the synthetic DNA sequence alignment. For the heuristic training scheme (left subfigure) the overall pattern of the posterior probabilities is correct, showing an increase for state $S_t = 2$ in the region $200 < t < 400$, and an increase for state $S_t = 3$ in the region $600 < t < 800$. However, the signals are very noisy, and an automatic classification based on the mode of the posterior probability would incur a high proportion of erroneously predicted topology changes. This shortcoming is significantly improved as a result of using the maximum likelihood scheme. The predicted state transitions coincide with the true breakpoints, and the tree topologies are predicted correctly. The posterior probabilities for the states, $P(S_t|\mathcal{D})$, are mostly close to zero or one. This indicates a high confidence in the prediction, which is reasonable: since the DNA sequence alignment results from the *simulation* of a recombination process, the transitions between topologies are, in fact, well defined. The estimated recombination parameter is $\nu = 0.992$. With four breakpoints in an alignment of length 1000 nucleotides, the correct value for the recombination parameter is $\nu = 0.996$, which deviates from the prediction by only 0.4%.

7.2. *Gene conversion in maize*

The prediction on the maize DNA sequence alignment is shown in Fig. 21. When using the heuristic parameter estimation method (left), the overall pattern of the graphs $P(S_t|\mathcal{D})$ captures the gene conversion event in that the final section shows a clear increase of the posterior probability for state $S_t = 3$. However, the signals are very noisy and unsuitable for an automatic detection of gene conversion without human intervention. The application of the maximum likelihood scheme leads to a clear improvement: a sharp transition from state $S_t = 1$ to state $S_t = 3$ is predicted in accordance with the gene conversion event found in [19].

7.3. *Recombination in Neisseria*

Figure 22 shows the prediction obtained on the *Neisseria* DNA sequence alignment. The heuristic training method (left) leads to a signal that is very noisy and only gives a vague indication of a topology change at the beginning of the alignment. Estimating the parameters with maximum likelihood leads to a considerable reduction in the noise. A topology change from state $S_t = 3$ to $S_t = 1$ with a breakpoint at site $t = 202$ is predicted, which is in accordance with the findings in [29]. Also, the second anomalous region between sites $t = 507$ and $t = 538$ is clearly detected in that the posterior probability for state 1, $P(S_t = 1|\mathcal{D})$, is significantly decreased, with sharp transitions at the sites predicted in [29]. However, while the HMM predicts a recombination event corresponding to a transition from state 1 into state 2, the findings in [29] suggest that this mosaic segment is more likely the result of rate variation than recombination. This will be discussed in more detail below.

8. Discussion

We have combined two probabilistic models for detecting interspecific recombination in DNA sequence alignments: (1) a taxon graph (phylogenetic tree) representing the relationships among the taxa, and (2) a site graph (HMM) representing which nucleotides interact in determining the tree topology. The parameters of the combined model can be estimated in a maximum likelihood sense with the EM algorithm, and this leads to a significant improvement on an older heuristic parameter estimation scheme. In fact, the simulation study carried out here suggests that recombinant regions can be accurately located, in agreement with the true location

(simulation study) or the location predicted in previous, independent work (maize actin genes, *Neisseria*).

Two limitations of the approach presented here, however, have to be discussed.

Each possible topology constitutes a separate hidden state of the HMM. Now recall, from Sec. 2.1, that for n taxa there are $(2n-5)!!$ different unrooted tree topologies. This implies that the number of states K increases super-exponentially with the number of taxa, which limits our algorithm to alignments of small numbers of taxa. In practical applications, the HMM method is therefore at best combined with a fast low-resolution preprocessing step that can analyze more taxa simultaneously. A useful approach is to conduct the initial search for recombination with split decomposition [2], a method that represents evolutionary relationships among sequences by a network if there are conflicting phylogenetic signals in the data. Split decomposition itself does not allow individual recombination events to be identified nor the statistical support for them to be assessed. It is, however, a useful preprocessing step in that a network that strongly deviates from a bifurcating tree is suggestive of recombination and gives hinds as to which sequences might belong to candidate recombinant strains. This can then be further investigated with the high-resolution method discussed in the present paper.

The second limitation is that the hidden states represent different tree topologies, but do not allow for different rates of evolution. However, if a region has evolved at a drastically different rate, employing a new state for modelling this region might increase the likelihood even though the new state itself — representing a different (wrong) topology — is ill-matched to

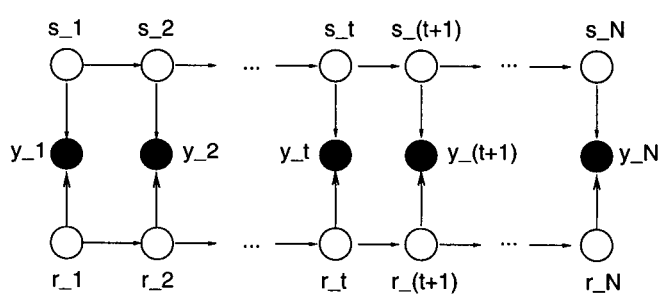

Fig. 23. **Factorial Hidden Markov Model.** In generalization of the standard HMM of Fig. 12, a factorial HMM has two separate families of hidden states: one represents different topologies (S_t), the other represents different evolutionary rates (r_t).

the data. Consequently, a *differently diverged* region might be erroneously classified as *recombinant*, which seems to have happened on the *Neisseria* sequence alignment, as discussed in the previous section. A way to redeem this deficiency is to employ a *factorial hidden Markov model*, shown in Fig. 23, and to introduce two separate hidden states: one representing different topologies, the other representing different evolutionary rates. This effectively combines the method of the present paper with the approach in [5]. A detailed investigation of this idea is the subject of future research.

References

[1] P. Baldi and P. Brunak, *Bioinformatics — The Machine Learning Approach* (MIT Press, Cambridge, MA, 1998).

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

[3] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Stat. Soc.* **B39**(1) (1977) 1–38.

[4] 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).

[5] J. Felsenstein and G. A. Churchill, A hidden Markov model approach to variation among sites in rate of evolution, *Molecular Biology Evolution* **13**(1) (1996) 93–104.

[6] J. Felsenstein, Evolution trees from DNA sequences: A maximum likelihood approach, *J. Molecular Evolution* **17** (1981) 368–376.

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

[8] N. Galtier and M. Gouy, Inferring patterns and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis, *Molecular Biology Evolution* **15**(7) (1998) 871–879.

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

[10] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, 3rd edn. (Oxford University Press, New York, 1985).

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

[12] D. Heckerman, A tutorial on learning with Bayesian networks, in *Learning in Graphical Models*, ed. M. I. Jordan, Adaptive Computation and Machine Learning (MIT Press, Cambridge, Massachusetts, 1999), pp. 301–354.

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

[14] D. Husmeier and F. Wright, Detection of recombination in DNA multiple alignments with hidden Markov models, *J. Computational Biology* **8**(4) (2001) 401–427.

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

[16] J. Maynard Smith, Analyzing the mosaic structure of genes, *J. Molecular Evolution* **34** (1992) 126–129.

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

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

[19] M. Moniz de Sa and G. Drouin, Phylogeny and substitution rates of angiosperm actin genes, *Molecular Biology Evolution* **13** (1996) 1198–1212.

[20] R. M. Neal and G. E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in *Learning in Graphical Models*, ed. M. I. Jordan (MIT Press, Cambridge, MA, 1999), pp. 355–368.

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

[22] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd edn. (McGraw-Hill, Singapore, 1991).

[23] D. Posada and K. A. Crandall, Selecting the best-fit model of nucleotide substitution, *Syst. Biology* **50**(4) (2001) 580–601.

[24] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE* **77**(2) (1989) 257–286.

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

[26] M. O. Salminen, J. K. Carr, D. S. Burke and F. E. McCutchan, Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning, *Aids Res. Human Retroviruses* **11**(11) (1995) 1423–1425.

[27] K. Strimmer and A. von Haeseler, Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies, *Molecular Biology Evolution* **13** (1996) 964–969.

[28] J. D. Thompson, D. G. Higgins and T. J. Gibson, CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice, *Nucleic Acids Res.* **22** (1994) 4673–4680.

[29] 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** (1992) 2135–2146.