Phylogenetic Hidden Markov Models

Adam Siepel and David Haussler

Center for Biomolecular Science and Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064, USA

Phylogenetic hidden Markov models, or phylo-HMMs, are probabilistic models that consider not only the way substitutions occur through evolutionary history at each site of a genome, but also the way this process changes from one site to the next. By treating molecular evolution as a combination of two Markov processes—one that operates in the dimension of *space* (along a genome) and one that operates in the dimension of *time* (along the branches of a phylogenetic tree)—these models allow aspects of both sequence structure and sequence evolution to be captured. Moreover, as we will discuss, they permit key computations to be performed exactly and efficiently. Phylo-HMMs allow evolutionary information to be brought to bear on a wide variety of problems of sequence "segmentation," such as gene prediction and the identification of conserved elements.

Phylo-HMMs were first proposed as a way of improving phylogenetic models that allow for variation among sites in the rate of substitution [8, 52]. Soon afterward, they were adapted for the problem of secondary structure prediction [10, 47], and some time later, for the detection of recombination events [19]. Recently there has been a revival of interest in these models [40, 42, 43, 44, 31], in connection with an explosion in the availability of comparative sequence data, and an accompanying surge of interest in comparative methods for the detection of functional elements [35, 3, 23, 46, 41]. There has been particular interest in applying phylo-HMMs to a multi-species version of the ab initio gene prediction problem [40, 43, 31].

In this chapter, phylo-HMMs are introduced, and examples are presented illustrating how they can be used both to identify regions of interest in multiply aligned sequences, and to improve the goodness of fit of ordinary phylogenetic models. In addition, we discuss how hidden Markov models (HMMs), phylogenetic models, and phylo-HMMs all can be considered special cases of general "graphical models," and how the algorithms that are used with these models can be considered special cases of more general algorithms. This chapter is written at a tutorial level, suitable for readers who are familiar with phylogenetic models but have had limited exposure to other kinds of graphical models.

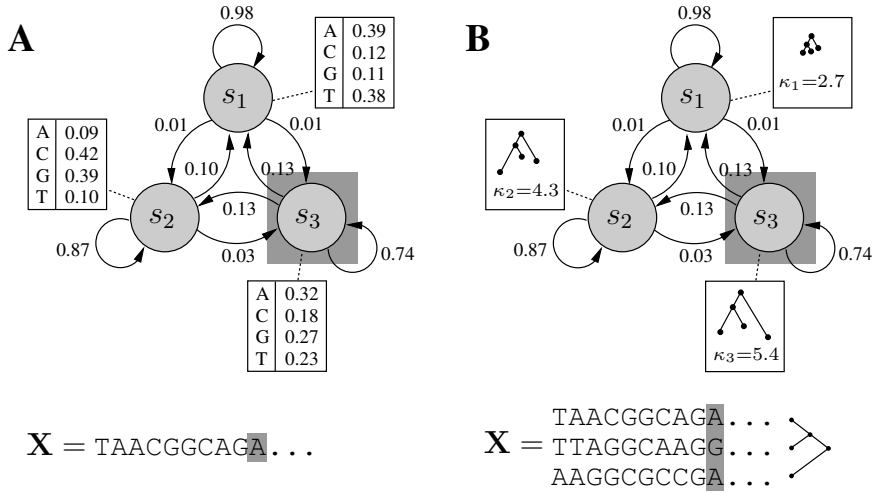

Fig. 1. (**A**) A 3-state single-sequence HMM, with a multinomial distribution associated with each state (boxed tables). A new state is visited at each time step, according to the indicated transition probabilities (numbers on arcs), and a new character is emitted, according to the probability distribution for that state. The shaded boxes indicate the current state and a newly emitted character, which is appended to the sequence **X**. In this example, one state has an A+T rich distribution (s_1), one has a G+C rich distribution (s_2), and one favors purines (s_3). (**B**) An analogous phylo-HMM. In this case, the multinomial distributions are replaced by phylogenetic models, and at each time step a new column in a multiple alignment **X** is emitted. The phylogenetic models include parameters describing the overall shape and size of the tree, as well as the background distribution for characters and the pattern of substitution. For simplicity, the tree parameters are represented graphically, and only one auxiliary parameter is shown.

1 Background

A phylo-HMM can be thought of as a machine that probabilistically generates a multiple alignment, column by column, such that each column is defined by a phylogenetic model. As with the single-sequence HMMs ordinarily used in biological sequence analysis [6], this machine probabilistically proceeds from one state to another[1], and at each time step it "emits" an observable object, which is drawn from the distribution associated with the current state (Figure 1). With phylo-HMMs, however, the distributions associated with states are no longer multinomial distributions over a set of characters (e.g., {A,C,G,T}), but are more complex distributions defined by phylogenetic models.

Phylogenetic models, as considered here, define a stochastic process of substitution that operates independently at each site in a genome (the question

[1] Throughout this chapter, it is assumed that the Markov chain for state transitions is discrete, first-order, and homogeneous.

of independence will be revisited below). In the assumed process, a character is first drawn at random from the background distribution and assigned to the root of the tree; character substitutions then occur randomly along the tree's branches, from root to leaves. The characters that remain at the leaves when the process has completed define an alignment column. Thus, a phylogenetic model induces a distribution over alignment columns having a correlation structure that reflects the phylogeny and substitution process (see [10]). The different phylogenetic models associated with the states of a phylo-HMM may reflect different overall rates of substitution (as in conserved and nonconserved regions), different patterns of substitution or background distributions (as in coding and noncoding regions), or even different tree topologies (as with recombination [19]).

Typically with HMMs, a sequence of observations (here denoted $\mathbf{X}$) is available to be analyzed, but the sequence of states (called the "path") by which the observations were generated is "hidden" (hence the name "hidden Markov model"). Efficient algorithms are available to compute the maximum-likelihood path, the posterior probability that any given state generated any given element of $\mathbf{X}$, and the total probability of $\mathbf{X}$ considering all possible paths (the likelihood of the model). The usefulness of HMMs in general, and phylo-HMMs in particular, is in large part a consequence of the fact that these computations can be performed exactly and efficiently. In this chapter, three examples of applications of phylo-HMMs will be presented that parallel these three types of computation—prediction based on the maximum-likelihood path (example 1), prediction based on posterior probabilities (example 2), and improved goodness of fit, as evidenced by model likelihood (example 3). Finally, it will be shown how these algorithms may be considered special cases of more general algorithms, by regarding phylo-HMMs as graphical models.

2 Formal Definition of a Phylo-HMM

Formally, we define phylo-HMM $\boldsymbol{\theta} = (S, \boldsymbol{\psi}, \mathbf{A}, \mathbf{b})$ to be a four-tuple, consisting of a set of states, $S = \{s_1, \ldots, s_M\}$, a set of associated phylogenetic models, $\boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_M\}$, a matrix of state-transition probabilities, $\mathbf{A} = \{a_{j,k}\}$ $(1 \leq j, k \leq M)$, and a vector of initial-state probabilities, $\mathbf{b} = (b_1, \ldots, b_M)$. In particular, $\boldsymbol{\psi}_j$ is the phylogenetic model associated with state s_j $(1 \leq j \leq M)$, $a_{j,k}$ $(1 \leq j, k \leq M)$ is the conditional probability of visiting state k at some site i given that state j is visited at site $i-1$, and b_j $(1 \leq j \leq M)$ is the probability that state j is visited first (thus, $\sum_k a_{j,k} = 1$ for all j, and $\sum_j b_j = 1$). Let $\mathbf{X}$ be the given alignment, consisting of L columns (sites) and n rows (one for each of n species), with the ith column denoted $\mathbf{X}_i$ $(1 \leq i \leq L)$.

Each phylogenetic model $\boldsymbol{\psi}_j$, in turn, consists of several components. For our purposes, a phylogenetic model $\boldsymbol{\psi}_j = (\mathbf{Q}_j, \boldsymbol{\pi}_j, \boldsymbol{\tau}_j, \boldsymbol{\beta}_j)$ is a four-tuple consisting of a substitution rate matrix $\mathbf{Q}_j$, a vector of background (or equilibrium) frequencies $\boldsymbol{\pi}_j$, a binary tree $\boldsymbol{\tau}_j$, and a set of branch lengths $\boldsymbol{\beta}_j$. The

model is defined with respect to an alphabet $\mathbf{\Sigma}$ (e.g., $\mathbf{\Sigma} = \{\text{A,C,G,T}\}$) whose size is denoted d. Generally, $\mathbf{Q}_j$ has dimension $d \times d$ and $\boldsymbol{\pi}$ has dimension d (but see example 3). The tree $\boldsymbol{\tau}_j$ has n leaves, corresponding to n present-day taxa. The elements of $\boldsymbol{\beta}_j$ are associated with the branches (edges) of the tree. It is assumed that all phylogenetic models in $\boldsymbol{\psi}$ are defined with respect to the same alphabet and number of species.

The probability that a column $\mathbf{X}_i$ is emitted by state s_j is simply the probability of $\mathbf{X}_i$ under the corresponding phylogenetic model, $P(\mathbf{X}_i|\boldsymbol{\psi}_j)$. This quantity can be computed efficiently by a recursive dynamic programming algorithm known as Felsenstein's "pruning" algorithm [7]. Felsenstein's algorithm requires conditional probabilities of substitution for all bases $a, b \in \mathbf{\Sigma}$ and branch lengths $t \in \boldsymbol{\beta}_j$. The probability of substitution of a base b for a base a along a branch of length t, denoted $P(b|a, t, \boldsymbol{\psi}_j)$, is based on a continuous-time Markov model of substitution, defined by the rate matrix $\mathbf{Q}_j$. In particular, for any given non-negative value t, the conditional probabilities $P(b|a, t, \boldsymbol{\psi}_j)$ for all $a, b \in \mathbf{\Sigma}$ are given by the $d \times d$ matrix $\mathbf{P}_j(t) = \exp(\mathbf{Q}_j t)$, where $\exp(\mathbf{Q}_j t) = \sum_{k=0}^{\infty} \frac{(\mathbf{Q}_j t)^k}{k!}$ [27]. $\mathbf{Q}_j$ can be parameterized in various more or less parsimonious ways [50]. For most of this chapter, we will assume the parameterization corresponding to the "HKY" model [12], which implies that $\mathbf{Q}_j$ has the form

$$\mathbf{Q}_j = \begin{pmatrix} - & \pi_{\text{C},j} & \kappa_j \pi_{\text{G},j} & \pi_{\text{T},j} \\ \pi_{\text{A},j} & - & \pi_{\text{G},j} & \kappa_j \pi_{\text{T},j} \\ \kappa_j \pi_{\text{A},j} & \pi_{\text{C},j} & - & \pi_{\text{T},j} \\ \pi_{\text{A},j} & \kappa_j \pi_{\text{C},j} & \pi_{\text{G},j} & - \end{pmatrix}, \tag{1}$$

where $\boldsymbol{\pi}_j = (\pi_{\text{A},j}, \pi_{\text{C},j}, \pi_{\text{G},j}, \pi_{\text{T},j})$, κ_j represents the transition/transversion rate ratio for model $\boldsymbol{\psi}_j$, and the $-$ symbols indicate quantities required to make each row sum to zero.

A "path" through the phylo-HMM is a sequence of states, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_L)$, such that $\phi_i \in \{1, \ldots, M\}$ for $1 \leq i \leq L$. The joint probability of a path and an alignment is[2]

$$P(\boldsymbol{\phi}, \mathbf{X}|\boldsymbol{\theta}) = b_{\phi_1} P(\mathbf{X}_1|\boldsymbol{\psi}_{\phi_1}) \prod_{i=2}^{L} a_{\phi_{i-1},\phi_i} P(\mathbf{X}_i|\boldsymbol{\psi}_{\phi_i}). \tag{2}$$

The likelihood is given by the sum over all paths, $P(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{\phi}} P(\boldsymbol{\phi}, \mathbf{X}|\boldsymbol{\theta})$, and the maximum-likelihood path is $\hat{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi}} P(\boldsymbol{\phi}, \mathbf{X}|\boldsymbol{\theta})$. These quantities can be computed efficiently using two closely related dynamic-programming algorithms known as the "forward" and Viterbi algorithms, respectively. The posterior probability that observation $\mathbf{X}_i$ was produced by state s_j, denoted $P(\phi_i = j|\mathbf{X}, \boldsymbol{\theta})$, can be computed for all i and j by combining the forward algorithm with a complementary "backward" algorithm, in a "forward-backward" procedure. Details can be found in [6].

[2] For simplicity, transitions to an "end" state are omitted here.

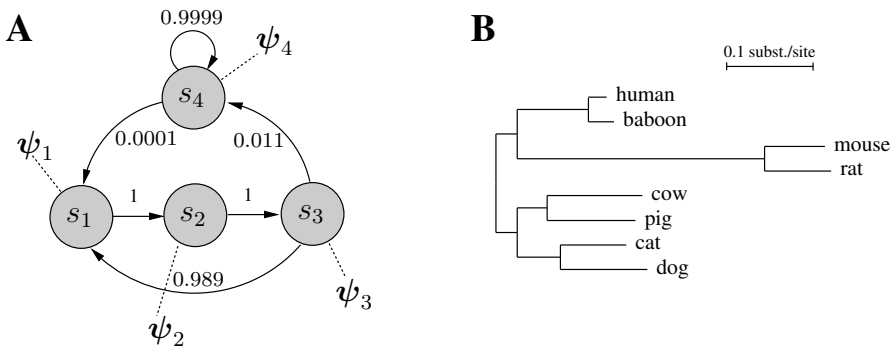

Fig. 2. (**A**) A 4-state phylo-HMM for gene finding. States s_1, s_2, and s_3 represent the three codon positions and state s_4 represents noncoding sites. The associated phylogenetic models $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_4$ capture characteristic properties of the different types of sites—e.g., the higher average rate of substitution, and the greater transition/transversion ratio, in noncoding and 3rd-codon-position sites than in 1st- and 2nd-codon-position sites. (**B**) The eight mammals and phylogeny assumed for the simulation, with branch lengths drawn in the proportions of the noncoding model ($\boldsymbol{\psi}_4$). Subsets of species were selected to maximize the sum of the branch lengths of the induced subtree—e.g., rat and dog for $n = 2$, and rat, dog, and cow for $n = 3$.

Example 1. A toy gene finder

This example is meant to demonstrate, in principle, how a phylo-HMM can be used for gene finding. Consider a simple 4-state phylo-HMM, with states for the three codon positions and noncoding sites (Fig. 2A). The problem is to identify the genes in a synthetic data set based on this model, using nothing but the aligned sequence data and the model (this is a multiple-sequence version of the ab initio gene prediction problem). For simplicity, we assume the model parameters $\boldsymbol{\theta}$ are given, along with the data set $\mathbf{X}$. In practice, the parameters have been set to reasonable values for a phylogeny of $n = 8$ mammals (Fig. 2B)[3], and the data set has been generated according to these values. The state path was recorded during the generation of the data, so that it could be used to evaluate the accuracy of predictions. The synthetic data set consists of $L = 100,000$ sites and 74 genes.

The Viterbi algorithm can be used for prediction of genes in this data set in a straightforward way. For every site i ($1 \leq i \leq L$) and state j ($1 \leq j \leq M$), the emission probability $P(\mathbf{X}_i|\boldsymbol{\psi}_j)$ is computed using Felsenstein's algorithm. These $L \times M$ values, together with the state-transition probabilities $\mathbf{A}$ and initial-state probabilities $\mathbf{b}$, are sufficient to define the joint probability $P(\boldsymbol{\phi}, \mathbf{X}|\boldsymbol{\psi})$ for any path $\boldsymbol{\phi}$, and can be simply plugged into the standard

[3] Parameter estimates from [44] were used for the phylogenetic models, and the state-transition probabilities were approximately based on estimates from [43] (the probability from s_4 to s_1 was inflated so that genes would not be too sparse). A uniform distribution was assumed for initial-state probabilities.

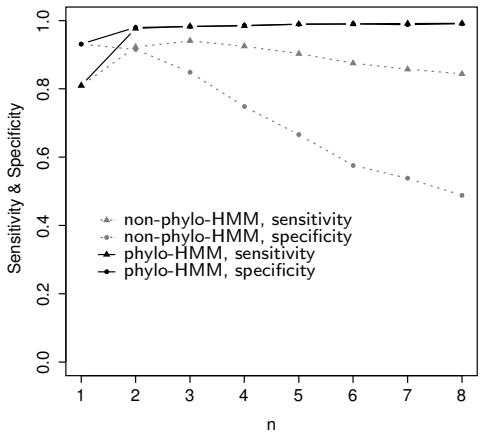

Fig. 3. Nucleotide-level sensitivity and specificity for the phylo- and non-phylo-HMMs on the simulated data set of example 1. Results are shown for $n = 1, \ldots, 8$ species.

Viterbi algorithm to obtain a maximum-likelihood path, $\hat{\phi}$. This predicted path, in turn, defines a set of predicted genes.

To evaluate the effect on prediction accuracy of the number of species in the data set, subsets of $n = 1, \ldots, 8$ sequences were selected from the full alignment (Fig. 2B), and a separate set of predictions was produced for each subset. Predictions were also produced with an alternative model, in which emission probabilities were based on the assumption that all characters in a column were independently drawn from the background (equilibrium) distribution of each state—in other words, the correlation structure implied by the phylogeny was ignored. This model, which will be called the "non-phylo-HMM," allows the importance of the phylogeny in the phylo-HMM to be assessed.

The nucleotide-level sensitivity (portion correctly predicted of sites actually in genes) and specificity (portion correct of sites predicted to be in genes) for both models are shown in Fig. 3, as the number of species increases from $n = 1$ to $n = 8$. The two models are identical for $n = 1$ (where there is no phylogeny to consider), but as the number of species increases from $n = 2, \ldots, 8$, the performance of the phylo-HMM rapidly improves, with about 98% sensitivity and specificity achieved by $n = 2$, and 99% sensitivity and specificity achieved by $n = 5$. The non-phylo-HMM, on the other hand, appears to improve slightly then decline, in both sensitivity and specificity.[4] The phylo-HMM is able to capitalize on differences in branch lengths and substitution

[4] It might be expected that the prediction accuracy of the non-phylo-HMM would simple fail to improve as rapidly as that of the phylo-HMM, rather than declining. The reason for the decline seems to be that the erroneous assumption of independence causes random fluctuations in base composition to appear more

patterns, while the non-phylo-HMM has to rely completely on more subtle differences in base composition.

This example is obviously a gross simplification of the real gene prediction problem: here, the model used for prediction exactly matches the model used to generate the data, while in the real problem, the model for prediction tends to fit the data in a much more approximate way (see Discussion). Even if slightly contrived, however, this example should help to illustrate how the information encoded in substitution rates and patterns can be exploited in problems of segmentation, such as gene prediction. □

Example 2. Identification of highly conserved regions

Our second example is concerned with a phylo-HMM in which states correspond to "rate categories"—classes of sites assumed to differ only in overall rate of substitution—rather than "functional categories," as in the previous example. The problem is to identify highly conserved genomic regions in a set of multiply aligned sequences. Such regions are likely to be functionally important, and hence, their identification has become a subject of considerable interest in comparative genomics; see Margulies et al. [30] for a recent review and a comprehensive discussion. In this example, we will use a phylo-HMM to identify conserved regions in a subset of the data set analyzed by Margulies et al. It will be shown that a phylo-HMM can be used to obtain results comparable to theirs, and has certain potential advantages over their methods.

A phylo-HMM is assumed like the one proposed by Felsenstein and Churchill [8], with k states corresponding to k rate categories, and state transitions defined by a single "autocorrelation" parameter λ (Fig. 4; a similar model, but with a more complex parameterization of transition probabilities, was proposed by Yang [52]). Regions of the alignment that are likely to have been generated by the "slowest" rate categories will be considered putative "Multi-species Conserved Sequences" (MCSs) [30]. Specifically, we will look at sites i for which the posterior probability $P(\phi_i = 1|\mathbf{X}, \boldsymbol{\theta})$ is high, assuming state s_1 has the smallest rate constant. Posterior probabilities will be computed using the forward-backward algorithm. As in example 1, the $L \times k$ table of emission probabilities—i.e., $P(\mathbf{X}_i|\boldsymbol{\psi}_j)$ for every site i $(1 \leq i \leq L)$ and state j $(1 \leq j \leq k)$—together with the state-transition and initial-state probabilities (parameters $\mathbf{A}$ and $\mathbf{b}$ of the phylo-HMM), can be plugged into the standard forward-backward algorithm for HMMs. In other words, once the emission probabilities are computed, the phylogenetic models can be ignored, and the phylo-HMM can be treated like an ordinary HMM. Note that inferences about the evolutionary rate at each site could alternatively be based on the Viterbi path. We have opted to use posterior probabilities instead, partly for illustration, and partly because they can be conveniently interpreted as a continuous-valued "conservation score" that can be plotted along the genome

significant than they really are. These fluctuations are "explained" by changes in state, resulting in errors in the inferred path and a decline in accuracy.

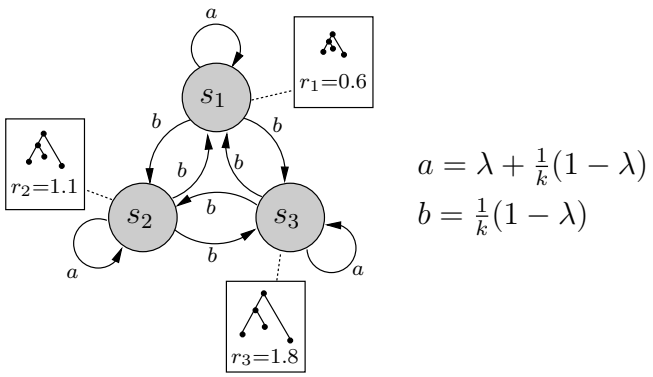

Fig. 4. State-transition diagram for the autocorrelated rate-variation model of Felsenstein and Churchill [8], with $k = 3$ rate categories and a uniform stationary distribution. The autocorrelation parameter λ defines all transition probabilities, as shown. It takes values between 0 and 1, and describes the degree to which the evolutionary rates at adjacent sites tend to be similar. The values r_1, r_2, and r_3 are applied as scaling constants to the branch lengths of a phylogenetic model; all parameters other than branch lengths are left unchanged. In our case, these "rate constants," as well as λ, are estimated (approximately) from the data (see [42]).

(see below). With this model, the posterior probabilities also tend to be more robust than the Viterbi path, which is highly sensitive to λ.

The data set consists of about 1.8 Mb of human sequence from chromosome 7 and homologous sequence from 8 other eutherian mammals [46] (we consider only the 9 mammals of the 12 species analyzed in [30]). The species and phylogeny are as shown in Fig. 2B, except that in this case, chimp is also included, and appears in the phylogeny as a sister taxon to human. Assuming the HKY substitution model and $k = 10$ states, we fitted a phylo-HMM to this alignment, obtaining an estimate of $\hat{\lambda} = 0.94$. Using these parameter estimates, we then computed the posterior probability of each state at each site. The posterior probabilities for s_1 in a selected region of the alignment are shown in Fig. 5, along with the conservation scores developed by Margulies et al. The known exons in this region all coincide with regions of high posterior probability, as do several conserved intronic features identified by Margulies et al. [30].

A detailed comparison of results is not possible here, but we note that the posterior probabilities based on the phylo-HMM are qualitatively very similar to the binomial- and parsimony-based conservation scores of Margulies et al. [30]. In addition, the phylo-HMM may have certain advantages as a framework for addressing this problem. For example, it requires no sliding window of fixed size, and as a result, is capable of identifying both very short highly conserved sequences, and long not-so-conserved sequences. In addition, it can be used with any phylogenetic model, including, e.g., ones that allow for non-

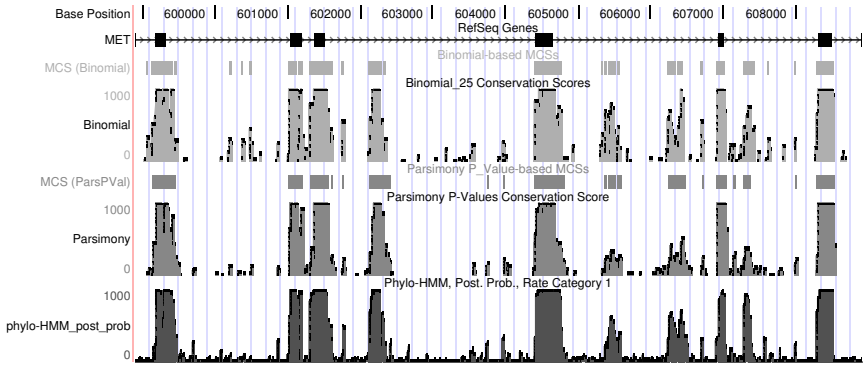

Fig. 5. A screen shot from the UCSC Genome Browser [24] showing a selected region of the data set of example 2, including several exons of the *MET* gene (black boxes at top). The binomial-based (light gray) and parsimony-based (medium gray) conservation scores of Margulies et al. [30] are shown as tracks in the browser, as are the posterior probabilities ($\times 1000$) of state s_1 in the phylo-HMM (dark gray). Plots similar to this one, showing phylo-HMM-based conservation scores across the whole human genome, can be viewed online at http://genome.ucsc.edu.

homogeneities in the substitution process or context-dependent substitution (see example 3); it extends naturally to the case in which different functional categories of sites, as well as rate categories, are considered [42]; and it could be adapted to model properties such as the length distributions of MCSs (e.g., using techniques from gene finding). □

3 Higher Order Markov Models for Emissions

It is common with (single-sequence) gene-finding HMMs to condition the emission probability of each observation x_i on the observations that immediately precede it in the sequence, e.g., x_{i-2} and x_{i-1}. By taking into consideration the "context" for each observation, emission probabilities become more informative, and the HMM can discriminate more effectively between different classes of observations. For example, in a 3rd-codon-position state, the emission of a base x_i = "A" might have a fairly high probability if the previous two bases are x_{i-2} = "G" and x_{i-1} ="A" (GAA = Glu), but should have zero probability if the previous two bases are x_{i-2} = "T" and x_{i-1} = "A" (TAA = Stop).

Considering the N observations preceding each x_i corresponds to using an Nth order Markov model for emissions. (Note that such a model does not imply an Nth order Markov chain for state transitions; indeed, things are kept simpler, and the model remains mathematically valid, if state transitions continue to be described by a 1st order Markov chain.) An Nth order model for emissions is typically parameterized in terms of $(N+1)$-tuples of observations,

and conditional probabilities are computed as

$$P(x_i|x_{i-N},\ldots,x_{i-1}) = \frac{P(x_{i-N},\ldots,x_{i-1},x_i)}{\sum_y P(x_{i-N},\ldots,x_{i-1},y)}, \tag{3}$$

with the numerator being the probability of the $(N+1)$-tuple $(x_{i-N},\ldots,x_i)$, and the sum in the denominator being over all possible observations y that could appear in place of x_i.

An Nth order Markov model for emissions can be incorporated into a phylo-HMM in essentially the same way. In this case, a whole alignment column $\mathbf{X}_i$ is considered in place of each single base x_i. Because we will primarily be concerned below with tuple size, let us also redefine N and speak of $(N-1)$st order Markov models and N-tuples of observations, instead of Nth order Markov models and $(N+1)$-tuples of observations. With these changes, equation 3 can be rewritten as

$$P(\mathbf{X}_i|\mathbf{X}_{i-N+1},\ldots,\mathbf{X}_{i-1}) = \frac{P(\mathbf{X}_{i-N+1},\ldots,\mathbf{X}_{i-1},\mathbf{X}_i)}{\sum_{\mathbf{Y}} P(\mathbf{X}_{i-N+1},\ldots,\mathbf{X}_{i-1},\mathbf{Y})}. \tag{4}$$

Notice that the sum in the denominator is now over all possible alignment columns $\mathbf{Y}$, and has d^n terms, where d is the size of the alphabet ($d = |\mathbf{\Sigma}|$) and n is the number of rows (species) in the alignment. To compute the quantity in the numerator of equation 4, we replace an ordinary phylogenetic model, defined with respect to an alphabet $\mathbf{\Sigma}$, with what we will call an "Nth order" phylogenetic model, defined with respect to $\mathbf{\Sigma}^N$, the alphabet of N-tuples of characters from $\mathbf{\Sigma}$.[5] (The new rate matrix and vector of equilibrium frequencies will have dimensions $d^N \times d^N$ and d^N, respectively.) The N-tuple of columns in the numerator is reinterpreted as a column of N-tuples, and its probability is computed with Felsenstein's pruning algorithm, using the Nth order phylogenetic model. The sum in the denominator can no longer be evaluated directly, but it can be computed efficiently by dynamic programming, using a slight adaptation of Felsenstein's algorithm [44, 42]. This new algorithm differs from the original only in its initialization strategy. Thus, the conditional probability $P(\mathbf{X}_i|\mathbf{X}_{i-N+1},\ldots,\mathbf{X}_{i-1})$ can be computed with an Nth order phylogenetic model and two passes through Felsenstein's algorithm, one for the numerator and one for the denominator of equation 4. This procedure is feasible only for small N, so far for $N \leq 3$.

[5] Note that the "order" of a phylogenetic model is given by the size of the tuples considered, and is not equal to the order of the Markov model for emissions. Here, Nth order phylogenetic models are used to define an $(N-1)$st order Markov model.

Once the conditional emission probabilities of equation 4 are available, they can be substituted directly into equation 2. For example, in the case of $N = 3$, equation 2 can be rewritten as

$$P(\boldsymbol{\phi}, \mathbf{X}|\boldsymbol{\theta}) = b_{\phi_1} P(\mathbf{X}_1|\boldsymbol{\psi}_{\phi_1}) a_{\phi_1,\phi_2} P(\mathbf{X}_2|\mathbf{X}_1, \boldsymbol{\psi}_{\phi_2}) \times \prod_{i=3}^{L} a_{\phi_{i-1},\phi_i} P(\mathbf{X}_i|\mathbf{X}_{i-2}, \mathbf{X}_{i-1}, \boldsymbol{\psi}_{\phi_i}). \quad (5)$$

The forward, Viterbi, and forward-backward algorithms are unaffected by the use of a higher-order Markov model for emissions.

It is important to note that this strategy for incorporating higher order Markov models into a phylo-HMM allows "context" to be considered in the nucleotide substitution process, as well as in the equilibrium frequencies of bases. Nth order phylogenetic models describe the joint substitution probabilities of N-tuples of nucleotides. As a result, the conditional probabilities of equation 4 may reflect various important context- or neighbor-dependencies in the substitution process, such as the tendency for synonymous substitutions to occur at a higher rate than nonsynonymous substitutions in coding regions, or the tendency for a high rate of C→T transitions in CpG dinucleotides. Equations 4 and 5, as will be shown in example 3, essentially provide a way of "stringing together" context-dependent phylogenetic models, so that context dependencies can be considered between every adjacent pair of columns in an alignment.

Example 3. Modeling context-dependent substitution

In this example, we will look at how goodness of fit is affected by increasing the order N of a phylogenetic model, and by allowing for Markov dependence between sites (as in equation 5). We will consider the goodness of fit of various independent-site ($N = 1$) and context-dependent ($N > 1$) phylogenetic models, with respect to about 160,000 sites in aligned noncoding DNA from 9 mammalian species. The results presented here are taken from [44]. (The full paper should be consulted for complete details.)

For convenience, let us call the class of phylo-HMMs described by equations 4 and 5 "Markov-dependent" models, because they allow for Markov dependence of columns in the alignment. As will be seen below, these models are actually only approximations of models that properly allow for Markov dependence across sites in the substitution process. Regardless, these Markov-dependent models are valid probability models (the probabilities of all alignments of a given size sum to one), so it is fair to evaluate goodness of fit based on model likelihoods. The way in which these models are approximate is discussed in detail in Section 7 and the Appendix.

In this example, there are no functional or rate categories to consider. We assume that the HMM has only a single state, so nothing is actually "hidden"—only one path is possible, and the model reduces to a Markov chain.

As a result, Equation 5 becomes

$$P(\boldsymbol{\phi}, \mathbf{X}|\boldsymbol{\theta}) = P(\mathbf{X}_1|\boldsymbol{\psi}_1)P(\mathbf{X}_2|\mathbf{X}_1, \boldsymbol{\psi}_1)\prod_{i=3}^{L} P(\mathbf{X}_i|\mathbf{X}_{i-2}, \mathbf{X}_{i-1}, \boldsymbol{\psi}_1). \quad (6)$$

This simplification allows us to focus on the impact of higher-order Markov models, and to avoid issues related to the HMM structure. Keep in mind, however, that higher-order Markov models can be used with a nontrivial HMM as easily as with this trivial one.

In [44], various models were fitted to the data set of 160,000 noncoding sites, and their likelihoods were compared. The models differed in the type of phylogenetic model used (its order N and the parameterization of its rate matrix), and whether N-tuples of columns were assumed independent or Markov-dependence was allowed. We will focus here on four types of phylogenetic models: the HKY and UNR 1st order models, the U2S 2nd order model, and the U3S 3rd order model. The HKY model, introduced in Section 2, is treated as a baseline. The UNR, or "unrestricted," model has a separate free parameter for every nondiagonal element of the rate matrix, and is the most general model possible for single nucleotide substitution (see, e.g., [51]). The U2S and U3S models are fully general 2nd and 3rd order models, respectively, except that they assume strand symmetry (so that, e.g., the rate at which AG changes to AC is the same as the rate at which CT changes to GT), and like most codon models [11], they prohibit instantaneous substitutions of more than one nucleotide. They have 48 and 288 rate-matrix parameters, respectively. We will consider two cases for each phylogenetic model: an "independent tuples" case, in which the data set was partitioned into N-tuples of columns, which were considered independent; and a Markov-dependent case, in which N-tuples were allowed to overlap, and likelihoods were computed with equations 4 and 6. Note that, with 1st order models, the independent-tuples and Markov-dependent cases are identical.

Fig. 6A shows the log likelihoods of the UNR, U2S, and U3S phylogenetic models, with and without Markov dependence, relative to the log likelihood of the HKY model. Even when N-tuples are considered independent, context-dependent models (here, U2S and U3S) produce a striking improvement in likelihood—a far larger increase than is obtained by replacing even a fairly parsimonious 1st order model (HKY) with a fully general one (UNR). When Markov dependence between sites is introduced, another large improvement occurs. This improvement appears to be largely a consequence of the fact that, with Markov dependence, every boundary between adjacent sites is considered, while with independent tuples, only every other (U2S) or every third (U3S) such boundary is considered. Notice that, even with Markov dependence, goodness of fit improves significantly when a 2nd order model (U2S) is replaced with a 3rd order model (U3S). This is probably partly because of direct context effects that extend beyond the nearest neighbors of each base, and partly because the 3rd order model does a better job than the

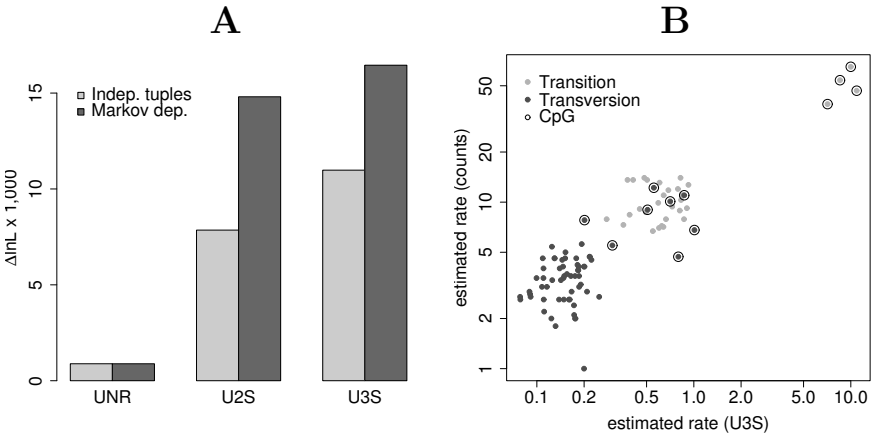

Fig. 6. (**A**) Log likelihoods of the UNR, U2S, and U3S phylogenetic models, with and without Markov dependence between sites, relative to the log likelihood of the HKY model. Results are for an alignment of 9 species and approximately 160,000 sites of noncoding data, as described in [44]. (**B**) Parameter estimates of substitution rates for the U3S model vs. estimates based on counts from aligned human genes and pseudogenes [15]. The rates cluster into three groups: transversions, transitions, and CpG transitions (CpG transversions cluster with non-CpG transitions). In general, the two sets of estimates agree fairly well, considering the differences in methods and data sets (see [44] for a detailed discussion).

2nd order model of accounting for indirect context effects—i.e., it provides a better approximation of a proper process-based model of context-dependent substitution (see below).

The observed improvements remain essentially unchanged when a measurement is used that considers the different numbers of parameters in the models and the size of the data set (the Bayesian information criterion) and in cross-validation experiments [44]. Thus, the apparent improvement in goodness of fit is not an artifact of the number of parameters in the models.

The U2S and U3S models allow context-dependent substitution rates to be estimated with full consideration of the phylogeny and allowance for multiple substitutions per site, unlike simpler "counting" methods for estimating context-dependent substitution rates [15]. Parameter estimates indicate wide variation in rates, spanning a 200-fold range, and in particular, pronounced CpG effects (Fig. 6B).

Coding regions can be modeled using a simple 3-state phylo-HMM, with a separate 3rd order phylogenetic model for each codon position. Thus, the state corresponding to the 3rd codon position considers columns of aligned codons, like an ordinary codon model, but the other two states consider columns of nucleotide triples that are out of frame, and consequently, these states can capture context effects that cross codon boundaries. Such a model improves

substantially on ordinary codon models, indicating that context effects that cross codon boundaries are important [44] (see also [39]). □

4 Phylogenetic Models, HMMs, and Phylo-HMMs as Graphical Models

In recent years, probabilistic models originally developed in various research communities have been unified under the heading of "graphical models." Graphical models provide an intuitively appealing framework for constructing and understanding probabilistic models, and at the same time, allow for rigorous analysis, in very general statistical and graph-theoretic terms, of algorithms for inference and learning. Many familiar classes of models fit naturally into the graphical models framework, including HMMs and phylogenetic models, as well as mixture models and hierarchical Bayesian models. A phylo-HMM can be seen as a graphical model whose structure is a hybrid of the graphical models for HMMs and phylogenetic models (Fig. 7). Viewing phylo-HMMs as graphical models helps to provide insight about why they permit efficient inference, and why this property may be sacrificed when assumptions such as site independence are relaxed. Our discussion of graphical models will necessarily be brief; other tutorials should be consulted for a more complete introduction to the field (e.g., [5, 13, 22]).

In graphical models, random variables are represented by nodes in a graph, and dependencies between variables are represented by edges (Fig. 7)[6]. Let X be the set of random variables represented by a graph with nodes (vertices) V and edges E, such that X_v is the variable associated with $v \in V$. In addition, let X_C be the subset of variables associated with $C \subseteq V$, and let lower-case letters indicate (sets of) instances of variables, e.g., x_v, x_C, and x. Graphical models can be defined in terms of directed or undirected graphs, and accordingly, are called directed or undirected models; here we will focus on the directed case, which for our purposes is simpler to describe. In a directed model, the edges of the graph correspond to local conditional probability distributions, and the joint probability of a set of instances x is a product of the conditional probabilities of nodes given their parents,

$$P(x) = \prod_{v \in V} P(x_v | x_{\mathcal{P}_v}), \tag{7}$$

where $\mathcal{P}_v$ denotes the set of parents of node v and $P(x_v|x_{\mathcal{P}_v})$ is the local conditional probability associated with x_v. It should not be too hard to see, looking at Fig. 7, that equation 7 generalizes the joint probability of a sequence and a particular path in the case of an HMM, and the joint probability of an

[6] The brief introduction to graphical models provided here roughly follows the more detailed tutorial of Jordan and Weiss [22].

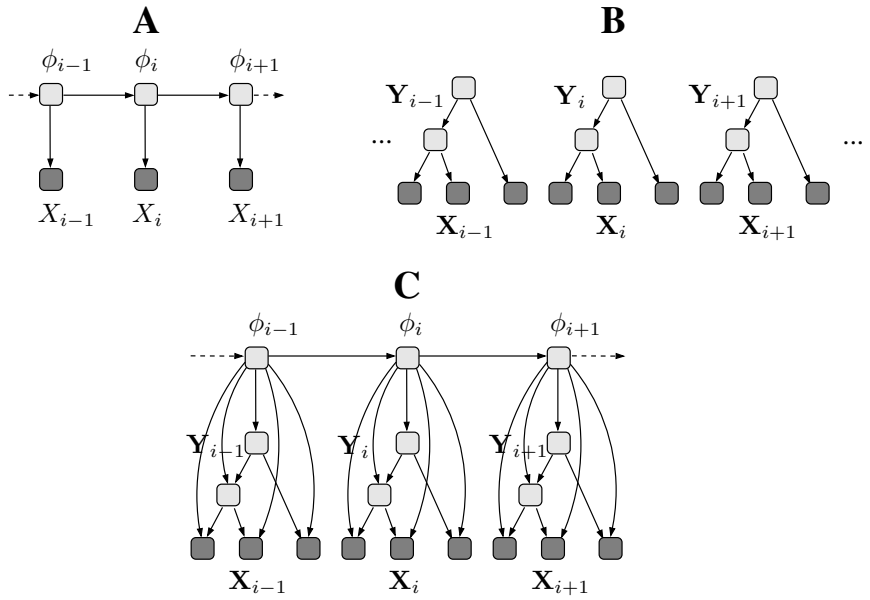

Fig. 7. Graphical model representations of (**A**) an HMM, (**B**) a phylogenetic model, and (**C**) a phylo-HMM. In each case, nodes correspond one-to-one with random variables; shaded nodes represent observed variables and unshaded nodes represent unobserved (latent) variables. These are directed graphical models, based on directed acyclic graphs (sometimes called Bayesian networks). The edges between nodes correspond to local conditional probability distributions, and can be thought of as implying dependencies between variables. (More precisely, the set of all edges defines a set of conditional independence assertions about the variables.) In (**A**), each X_i represents an observation in the sequence and each ϕ_i represents a state in the path. The conditional probability distribution for observation X_i given state ϕ_i is incorporated in the directed edge from ϕ_i to X_i, and the conditional probability distribution for state ϕ_i given state ϕ_{i-1} (i.e., of a transition from ϕ_{i-1} to ϕ_i) is incorporated in the directed edge from ϕ_{i-1} to ϕ_i. In (**B**), each set of nodes collectively labeled $\mathbf{X}_i$ represents an alignment column, and each set collectively labeled $\mathbf{Y}_i$ represents a set of ancestral bases. The conditional probabilities of nucleotide substitutions (based on the continuous-time Markov model) are incorporated in the directed edges from each parent node to its two children. In (**C**), conventions from (**A**) and (**B**) are combined.

alignment and a particular set of ancestral bases in the case of a phylogenetic model.

The general problem of *probabilistic inference* is to compute marginal probabilities from this joint distribution—probabilities of the form $P(x_U) = \sum_{x_W} P(x_U, x_W)$, where (U, W) is a partitioning of V. The likelihood is an example of such a marginal probability, with x_U being the observed data and X_W being the set of latent variables. When the likelihood of an HMM is com-

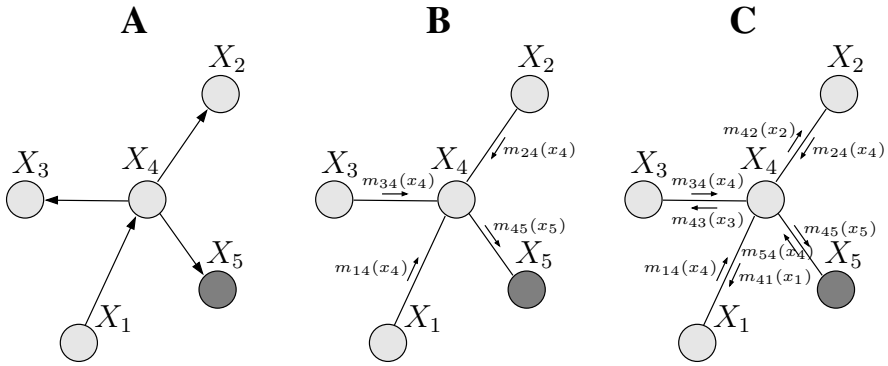

Fig. 8. (**A**) A directed graphical model whose nodes form an arbitrary tree. The marginal probability of an observed value of X_5 is desired. (**B**) The intermediate values of the elimination algorithm can be seen as "messages" that are passed from one node to another in the direction of X_5. (**C**) In the belief-propagation algorithm, all possible messages are generated simultaneously; the marginal probability of each node is a product of the incoming messages. (Based on Figure 1 of Jordan and Weiss [22].)

puted, x_U is the (observed) sequence and X_W is the (latent) path. With a phylogenetic model, the procedure is applied independently at each site, and x_U is an (observed) alignment column and X_W is a set of (latent) ancestral bases. Conditional probabilities of interest, such as the posterior probabilities of example 2, can be computed as quotients of marginal probabilities. For instance, suppose x_U is the observed data and X_w ($w \in W$) is a latent variable; then $P(x_w|x_U) = \frac{P(x_{U\cup\{w\}})}{P(x_U)}$.

Marginal probabilities can always be computed from the complete joint distribution by brute-force summation[7]. The problem is to keep these computations tractable as the number of random variables becomes large. It turns out that if a directed graphical model is a tree (or set of trees), as in Fig. 7A&B and Fig. 8, meaning that every node has at most one parent, then exact inference can be accomplished efficiently by dynamic programming. (As we will see, efficient exact inference is also possible in certain cases in which the directed graph is not a tree.)

The basic algorithm for computing marginal probabilities is known as "elimination," and is most easily described by example. Consider the graph of Fig. 8A, with $X = (X_1, X_2, X_3, X_4, X_5)$ and edges as depicted. The elimination algorithm takes advantage of the commutativity of sums and products, and reuse of intermediate computations, to reduce the computational complexity of a marginal summation.

[7] This discussion is restricted to discrete random variables, although it extends directly to the continuous case.

Algebraically, the algorithm proceeds as follows,

$$\begin{aligned}
P(x_5) &= \sum_{x_1,x_2,x_3,x_4} P(x_1,x_2,x_3,x_4,x_5) \\
&= \sum_{x_1}\sum_{x_2}\sum_{x_3}\sum_{x_4} P(x_1)P(x_2|x_4)P(x_3|x_4)P(x_4|x_1)P(x_5|x_4) \\
&= \sum_{x_4} P(x_5|x_4) \sum_{x_3} P(x_3|x_4) \sum_{x_2} P(x_2|x_4) \sum_{x_1} P(x_1)P(x_4|x_1) \\
&= \sum_{x_4} P(x_5|x_4) \sum_{x_3} P(x_3|x_4) \sum_{x_2} P(x_2|x_4) m_{14}(x_4) \\
&= \sum_{x_4} P(x_5|x_4) \sum_{x_3} P(x_3|x_4) m_{24}(x_4) m_{14}(x_4) \\
&= \sum_{x_4} P(x_5|x_4) m_{34}(x_4) m_{24}(x_4) m_{14}(x_4) \\
&= m_{45}(x_5), \qquad (8)
\end{aligned}$$

where the terms of the form $m_{ij}(x_j)$ denote the results of intermediate (nested) summations (each $m_{ij}(x_j)$ is the result of a sum over x_i and is a function of x_j). The algorithm can be described in graph-theoretic terms as a procedure that eliminates one node at a time from the graph until only the node corresponding to the desired marginal probability remains. From the algebraic description, many readers will recognize the similarity to Felsenstein's pruning algorithm [7]. Felsenstein's algorithm, it turns out, is an instance of the elimination algorithm—one of the earliest instances to be discovered. The forward algorithm is another instance of the elimination algorithm, as is the combined forward/Felsenstein algorithm that we used above to compute the likelihood of a phylo-HMM. The Viterbi algorithm is closely related to the elimination algorithm; it can be derived by noting that the "max" operator commutes with products, just as the summation operator does. Note that the elimination algorithm depends on a good "elimination ordering." An optimal ordering is difficult to find for arbitrary graphs, but can be determined easily for specific classes of models (as with HMMs, phylogenetic models, and phylo-HMMs).

Often, not just one, but many marginal probabilities are desired. The elimination algorithm can be extended to compute the marginal probabilities for all nodes in two passes across the graph, with conditional probabilities computed in a forward pass and marginals in a backward pass [29]. Typically, this procedure is described as "belief propagation" [37], with node elimination replaced by a "message-passing" metaphor (Fig. 8B&C). The belief-propagation (also called "sum-product") algorithm generalizes the forward-backward algorithm for HMMs and algorithms for phylogenetic models that compute marginal probabilities of ancestral bases [26].

We have focused on directed models, but undirected models are similar. Moreover, the undirected case turns out to be, in a sense, the more general one with respect to inference. In undirected models, the graph is viewed in

terms of cliques (maximal fully connected subgraphs) and a *potential function* (essentially, an unnormalized probability distribution) is associated with each clique. The joint probability of all variables (equation 7) is now a product over cliques, with a normalizing constant to ensure that $\sum_x P(x) = 1$. Directed graphs can be converted to undirected graphs by a process known as "moralization," wherein the arrowheads of the edges are removed and new edges are added between all parents of each node (the resulting graph is called the "moral" graph, because it requires that all parents are "married"). By explicitly creating a clique that includes each node and all of its parents, moralization ensures that all dependencies implied by the local conditional distributions of the directed graph are captured in the undirected graph.

The moral graph for a directed tree is simply an undirected tree (i.e., no new edges are added), and the belief propagation algorithm for this undirected tree is the same as that illustrated in Fig. 8. For undirected graphs that contain cycles, a generalization of the belief propagation algorithm, called the "junction-tree" algorithm, can be used. The junction-tree algorithm operates on a tree of cliques, rather than of nodes, and computes (unnormalized) marginal probabilities for cliques (marginal probabilities of nodes can be obtained afterwards). It requires an additional step, called "triangulation," in which new edges are added to the graph to represent certain implicit dependencies between nodes. A complete introduction to the junction-tree algorithm is not possible here (more details can be found in [5] and [22]). The key point for our purposes is that the computational complexity of the algorithm is exponential in the size of the largest clique. Thus, graphs with cycles can still be handled efficiently if their clique size is constrained to be small. It is for this reason that phylo-HMMs permit efficient inference; their (triangulated) moral graphs have cycles, but the maximum clique size turns out to be three[8]. When clique size is large, exact inference is intractable, and approximate methods are required. Some of the approximate methods in use include Monte Carlo algorithms and variational methods, which include mean field methods and "loopy" belief propagation (approximate methods are partially surveyed in [22]; see also [36, 53, 48, 49]).

With phylo-HMMs, the junction-tree algorithm allows computation not only of the posterior probability that each site was emitted by each state (as in example 2), but also of marginal posterior probabilities of ancestral bases considering all paths. In addition, the algorithm can be used to compute posterior expected values of interest, such as the expected number of substitutions per site, or the expected numbers of each type of substitution (A→C, A→G, etc.) along each branch of the tree (the sufficient statistics for parameter estimation by expectation maximization [9, 44]). Using the junction tree algorithm in the expectation step of an expectation-maximization

[8] In the case of a phylo-HMM, the parents of each node are already connected (Fig. 7C), so moralization amounts simply to removing the arrowheads from all edges in the graph. Moreover, it turns out that this graph is already triangulated.

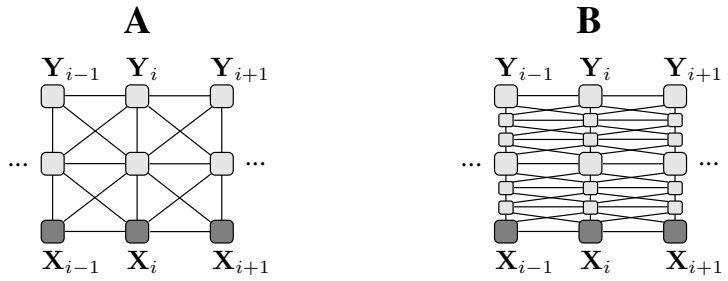

Fig. 9. (A) The lattice that results when context-dependent substitution is incorporated into a phylogenetic model, shown as an undirected graphical model. For clarity, only a single leaf node is shown for each site, with a chain of ancestral nodes leading to the root (the phylogeny can be imagined as going into and out of the page). Each node depends not only on its parent node in the phylogeny, but also on its parent's left and right neighbors in the alignment. (**B**) A version of the graph in (**A**) with intermediate nodes added to the branches of the tree. As more and more nodes are added, the branch lengths between them approach zero, and the model approaches a true "process-based" model of context-dependent substitution. In both (**A**) and (**B**), the untriangulated graph is shown; additional edges appear during triangulation, leading to prohibitively large clique sizes.

algorithm, it is possible to train a phylo-HMM (including its phylogenetic models) completely from unlabeled data. This technique could be used, for example, for de novo detection of binding-site motifs in aligned sequences.

Once the effect of cycles in graphical models is understood, it becomes clear that efficient exact inference will not be possible with models that accurately describe the *process* of context-dependent substitution, by allowing for dependencies between adjacent bases on all branches of the phylogenetic tree. Fig. 9A illustrates what happens to the graphical structure of a phylogenetic model when this kind of proper context-dependence is introduced. The additional edges in the graph lead to the formation of a kind of lattice of dependency, reminiscent of the classic Ising model from statistical mechanics (this case is like a two-dimensional Ising model, except that the branching structure of the phylogeny creates a branching structure of two-dimensional sheets, not shown in Fig. 9A). Unless the size of the lattice is constrained to be small, models of this kind are well-known to require approximate methods for inference.

Moreover, for context-dependent substitution to be modeled properly, it should be integrated into the continuous-time Markov model of substitution, so that context-effects can propagate indefinitely across sites as substitutions accumulate along each branch of the phylogeny. This behavior can be approximated by introducing intermediate nodes in the phylogeny, while keeping total branch lengths constant, as shown in Fig. 9B. As more and more nodes are introduced, the branch lengths between them will approach zero, and the model will approach the desired "process-based" model. Exact inference is

intractable for such models even in the case of two sequences and an unrooted tree, but Markov chain Monte Carlo (MCMC) methods have been applied in this special case [20, 38]. The stationary distribution of a related process has also been studied [2]. Extending such process-based models to full phylogenies appears difficult, even with MCMC. However, a model without intermediate nodes (as in Fig. 9A) has been studied by Jojic et al. [21], using variational methods for approximate inference. Jojic et al. have shown experimentally that this model can produce significantly higher likelihoods than the U2S version of the more approximate Markov-dependent model described in Section 3.

The model of Section 3 essentially works by defining a simple $(N-1)$st order Markov chain of alignment columns (observed variables), while ignoring the dependencies between the ancestral bases (latent variables) that are associated with overlapping N-tuples of columns. As a result, this model has no reasonable process-based interpretation. Nevertheless, it is a valid probability model that appears to fit the data well, and it allows for exact inference at modest computational cost [44]. The Markov-dependent model is compared to the model of Jojic et al. in more detail in the Appendix.

5 Discussion

Phylogenetic hidden Markov models are probabilistic models that describe molecular evolution as a combination of two Markov processes—one that operates in the dimension of *space* (along a genome) and one that operates in the dimension of *time* (along the branches of a tree). They combine HMMs and phylogenetic models, two of the most powerful and widely used classes of probabilistic models in biological sequence analysis. Phylo-HMMs often fit aligned DNA sequence data considerably better than models that treat all sites equally, or that fail to allow for correlations between sites. In addition, they are useful for identifying regions of interest in aligned sequences, such as genes or highly conserved regions.

Three examples have been presented to illustrate some of the ways in which phylo-HMMs may be used, and each one deserves additional comment. Applying phylo-HMMs to gene prediction (example 1) is a much harder problem than implied here, for several reasons. First, while coding and noncoding sites have quite different properties on average, both types of sites are heterogeneous mixtures, so that correctly classifying particular sequence segments can be difficult. For example, protein coding sites show higher average levels of evolutionary conservation than noncoding sites, but mammalian genomes do appear to have many islands of conservation in noncoding regions [4, 30], which can lead to false-positive predictions of exons [43]. Similarly, coding sites in mammalian genomes exhibit higher average G+C content than do noncoding sites, but base composition varies considerably in both kinds of sites from one genomic region to another, which can have the effect of confounding gene prediction software. Second, the gene finding problem ends up being largely

about identifying the boundaries of exons, as determined by splice sites, and phylo-HMMs are not necessarily the best tools for detecting these so-called "signals." Gene finders are often based on composite models, with specialized submodels for signal detection; a similar approach may be required for phylo-HMMs to be effective in gene prediction. A third problem is that a straightforward phylo-HMM like that of example 1 induces a geometric distribution of exon lengths, which is known to be incorrect. Some of these problems have been addressed with a "generalized" phylo-HMM that allows for arbitrary length distributions of exons, and also uses different sets of parameters for regions of different overall G+C content [31]. In other recent work, it has been shown that the prediction performance of a phylo-HMM-based exon predictor can be improved significantly by using context-dependent phylogenetic models, and by explicitly modeling both conserved noncoding regions and nucleotide insertions/deletions [43]. Additional challenges in multi-species gene prediction are also discussed in [43], stemming from lack of conservation of exon structure across species, and errors in the multiple alignment.

There are many possible ways of identifying conserved regions (example 2) and even quite different methods (e.g., ones that do and do not consider the phylogeny) tend to be fairly concordant in the regions they identify [45, 30]. Perhaps more difficult than proposing a method to identify conserved regions is confirming that it produces biologically useful results. Limited kinds of validation can be done computationally, but this is ultimately an experimental problem, and must be addressed in the laboratory. Most likely, phylo-HMMs of the kind described in example 2 will not produce dramatically different results from other methods, but as mentioned above, they provide a flexible framework in which to address the problem. It should be noted that, while the original papers introducing phylo-HMMs focused on improving the realism and goodness of fit of models allowing for rate variation [8, 52], they also showed that phylo-HMMs could be used to predict the evolutionary rate at each site.

Modeling context-dependent substitution is an active area of current research, and the Markov-dependent model described here (example 3) represents only one of several possible approaches to this problem. The approach of Jojic et al. [21], discussed at the end of Section 4, is another, and we are aware of work in progress on at least two other, completely different, methods. At this stage, it remains unclear which models and algorithms for inference will allow for the best compromise between computational efficiency and goodness of fit. It is likely that different approaches will turn out to be appropriate for different purposes.

Space has not allowed for a complete survey of the applications of phylo-HMMs. In particular, we have not discussed their use in the prediction of secondary structure [10, 47, 28] or the detection of recombination [19], nor have we touched on their use in a Bayesian setting [32, 18]. We also have not discussed the models similar in spirit to phylo-HMMs that have been applied to the problems of RNA secondary structure prediction [25] and multiple

alignment [34, 17, 16, 14]. It has been noted [40] that phylo-HMMs themselves could be used for multiple alignment, in a direct extension of the way pair HMMs are used for pairwise alignment [6]. Indeed, phylo-HMMs provide a natural framework for simultaneously addressing the multiple alignment and gene prediction problems, as has been done in the two-sequence case with pair HMMs [1, 33]. Another area in which phylo-HMMs may prove useful is homology searching. In principle, the profile HMMs that are commonly applied to this problem [6] could be adapted to use phylogenetic models instead of assuming independence of aligned sequences or relying on ad hoc weighting schemes.

Acknowledgments

We thank Nick Goldman, David Heckerman, and Michael Jordan for helpful discussions about context-dependent substitution, and Brian Lucena, Mathieu Blanchette, Robert Baertsch, and Michael Jordan for comments on the manuscript. A.S. is supported by an ARCS Foundation scholarship and NHGRI grant IP41HG02371, and D.H. is supported by the Howard Hughes Medical Institute.

Appendix

In this short Appendix, we will examine more closely how the Markov-dependent model for context-dependent substitution that was presented in Section 3 (example 3) compares with the graphical models of Section 4. We will concentrate on the model studied by Jojic et al. [21], which we will refer to as the "simple-lattice" model, in contrast to the full process-based model of Fig. 9B. The undirected graph for the simple-lattice model is shown in Fig. 10A, assuming a very small alignment of $n = 3$ sequences and $L = 3$ columns. (The complete graph is shown here, whereas in Fig. 9A only a subgraph was shown.) From Fig. 10A, it should be clear that the graph contains an $L \times 2$ lattice of nodes for each branch of the phylogeny.

The Markov-dependent model of Section 3 is a graphical model insofar as it is based on a Markov-chain of random variables, but it is quite different from the simple-lattice model. The Markov-dependent model actually operates at two levels, as illustrated in Fig. 10B. At one level (top of figure), a simple Markov chain of alignment columns is assumed, with each column being treated as an observed random variable. At another level (boxes at bottom of figure), the conditional probability of each column given the previous column is computed according to a phylogenetic model for pairs of columns. (Each of these phylogenetic models is a submodel of the model shown in Fig. 10A.) When conditional probabilities are computed according to these separate phylogenetic models, multiple versions of the random variables for ancestral bases

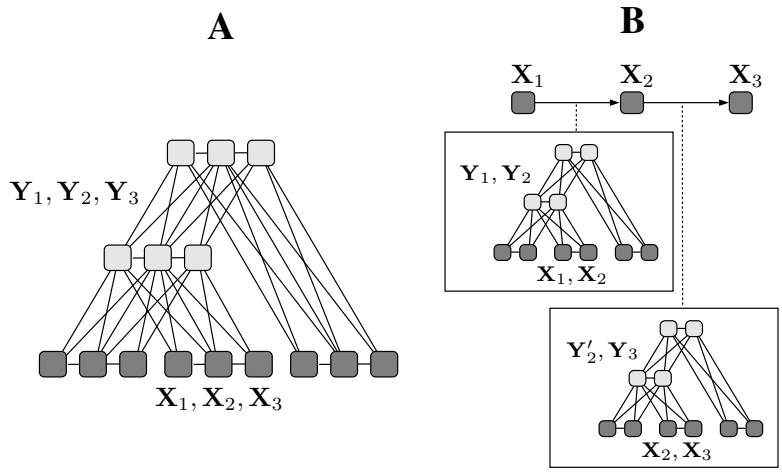

Fig. 10. (A) Undirected graph for the "simple-lattice" model of Fig. 9A, for an alignment of $L = 3$ sites and $n = 3$ species. Each node in the phylogeny is represented by a sequence of three nodes, corresponding to sites 1, 2, and 3, and each of these nodes is connected not only to its parent but to its parent's neighbors to the left and right. The shaded nodes together represent the three columns of the alignment, $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, and the unshaded nodes represent the corresponding sets of ancestral bases, $\mathbf{Y}_1$, $\mathbf{Y}_2$, and $\mathbf{Y}_3$. (**B**) An interpretation of the Markov-chain model of Section 3, applied to the same alignment (the case of $N = 2$ is illustrated). At one level (top), a Markov chain of alignment columns is assumed. At another level (bottom, inside boxes), the conditional probability of each column given the previous column is computed according to a phylogenetic model for pairs of sites.

are effectively introduced (e.g., $\mathbf{Y}_2$ and $\mathbf{Y}'_2$ in Fig. 10B). Moreover, these different versions are not required to be consistent. The effect of this modeling choice is to ignore (indirect) dependencies between latent variables that do not belong to the same "slice" of N columns, but at the same time, to permit exact likelihood computations and to capture what are probably the most important context effects.

By failing to "tie" together the ancestral nodes of these multiple phylogenetic models, the Markov-dependent model sacrifices any claim of accurately representing the process of context-dependent substitution. Nevertheless, it allows the major *consequences* of this process to be characterized empirically, in such a way that valid likelihoods can be extracted, as well as reasonable approximations of the conditional expectations of key quantities.

References

1. M. Alexandersson, S. Cawley, and L. Pachter. SLAM: Cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res.*, 13:496–502, 2003.
2. P. F. Arndt, C. B. Burge, and T. Hwa. DNA sequence evolution with neighbor-dependent mutation. In *Proc. 6th Int'l Conf. on Research in Computational Molecular Biology (RECOMB'02)*, pages 32–38, 2002.
3. D. Boffelli, J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter, and E. M. Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299:1391–1394, 2003.
4. F. Chiaromonte, R. J. Weber, K. M. Roskin, M. Diekhans, W. J. Kent, and D. Haussler. The share of human genomic DNA under selection estimated from human-mouse genomic alignments. In *Cold Spring Harbor Symp. Quant. Biol.*, volume 68, pages 245–254, 2003.
5. R. Cowell. Introduction to inference for Bayesian networks. In M. I. Jordan, editor, *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.
6. R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.
7. J. Felsenstein. Evolutionary trees from DNA sequences. *J. Mol. Evol.*, 17:368–376, 1981.
8. J. Felsenstein and G. A. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, 13:93–104, 1996.
9. N. Friedman, M. Ninio, I. Pe'er, and T. Pupko. A structural EM algorithm for phylogenetic inference. *J. Comp. Biol.*, 9:331–353, 2002.
10. N. Goldman, J. L. Thorne, and D. T. Jones. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.*, 263:196–208, 1996.
11. N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, 11:725–735, 1994.
12. M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22:160–174, 1985.
13. D. Heckerman. A tutorial on learning with Bayesian networks. In M. I. Jordan, editor, *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.
14. J. Hein, J. L. Jensen, and C. N. S. Pedersen. Recursions for statistical multiple alignment. *Proc. Natl. Acad. Sci. USA*, 100:14960–14965, 2003.
15. S. T. Hess, J. D. Blake, and R. D. Blake. Wide variations in neighbor-dependent substitution rates. *J. Mol. Biol.*, 236:1022–1033, 1994.
16. I. Holmes. Using guide trees to construct multiple-sequence evolutionary HMMs. *Bioinformatics*, 19(Suppl. 1):i147–i157, 2003.
17. I. Holmes and W. J. Bruno. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, 17:803–820, 2001.
18. D. Husmeier and G. McGuire. Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Mol. Biol. Evol.*, 20:315–337, 2003.
19. D. Husmeier and F. Wright. Detection of recombination in DNA multiple alignments with hidden Markov models. *J. Comp. Biol.*, 8:401–427, 2001.

20. J. L. Jensen and A.-M. K. Pedersen. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv. Appl. Prob*, 32:499–517, 2000.
21. V. Jojic, N. Jojic, C. Meek, D. Geiger, A. Siepel, D. Haussler, and D. Heckerman. Efficient approximations for learning phylogenetic HMM models from data. In *Proc. 12th Int'l Conf. on Intelligent Systems for Molecular Biology*, 2004. In press.
22. M. I. Jordan and Y. Weiss. Graphical models: probabilistic inference. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, 2nd edition, 2002.
23. M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423:241–254, 2003.
24. W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Res.*, 12:996–1006, 2002.
25. B. Knudsen and J. Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15:446–454, 1999.
26. J. M. Koshi and R. M. Goldstein. Probabilistic reconstruction of ancestral protein sequences. *J. Mol. Evol.*, 42:313–320, 1996.
27. P. Liò and N. Goldman. Models of molecular evolution and phylogeny. *Genome Res.*, 8:1233–1244, 1998.
28. P. Liò, N. Goldman, J. L. Thorne, and D. T. Jones. PASSML: Combining evolutionary inference and protein secondary structure prediction. *Bioinformatics*, 14:726–733, 1998.
29. B. Lucena. *Dynamic Programming, Tree-Width, and Computation on Graphical Models*. PhD thesis, Brown University, 2002.
30. E. H. Margulies, M. Blanchette, NISC Comparative Sequencing Program, D. Haussler, and E. D. Green. Identification and characterization of multi-species conserved sequences. *Genome Res.*, 13:2507–2518, 2003.
31. J. D. McAuliffe, L. Pachter, and M. I. Jordan. Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. *Bioinformatics*, 2004. In press.
32. G. McGuire, F. Wright, and M. J. Prentice. A Bayesian model for detecting past recombination events in DNA multiple alignments. *J. Comp. Biol.*, 7:159–170, 2000.
33. I. M. Meyer and R. Durbin. Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics*, 18:1309–1318, 2002.
34. G. J. Mitchison. A probabilistic treatment of phylogeny and sequence alignment. *J. Mol. Evol.*, 49:11–22, 1999.
35. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.
36. K. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief-propagation for approximate inference: An empirical study. In K. B. Laskey and H. Prade, editors, *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 467–476, San Mateo, CA, 1999. Morgan Kaufmann.
37. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

38. A.-M. K. Pedersen and J. L. Jensen. A dependent rates model and MCMC based methodology for the maximum likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.*, 18:763–776, 2001.
39. A.-M. K. Pedersen, C. Wiuf, and F. B. Christiansen. A codon-based model designed to describe lentiviral evolution. *Mol. Biol. Evol.*, 15:1069–1081, 1998.
40. J. S. Pedersen and J. Hein. Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics*, 19:219–227, 2003.
41. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway Rat yields insights into mammalian evolution. *Nature*, 428:493–521, 2004.
42. A. Siepel and D. Haussler. Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comp. Biol.*, 2004. In press.
43. A. Siepel and D. Haussler. Computational identification of evolutionarily conserved exons. In *Proc. 8th Int'l Conf. on Research in Computational Molecular Biology (RECOMB'04)*, pages 177–186, 2004.
44. A. Siepel and D. Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.*, 21:468–488, 2004.
45. N. Stojanovic, L. Florea, C. Riemer, D. Gumucio, J. Slightom, M. Goodman, W. Miller, and R. Hardison. Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Res.*, 27:3899–3910, 1999.
46. J. W. Thomas, J. W. Touchman, R. W. Blakesley, et al. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424:788–793, 2003.
47. J. L. Thorne, N. Goldman, and D. T. Jones. Combining protein evolution and secondary structure. *Mol. Biol. Evol.*, 13:666–673, 1996.
48. M. Wainwright, T. Jaakkola, and A. Willsky. Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Transactions on Information Theory*, 49:1120–1146, 2001.
49. M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley, 2003.
50. S. Whelan, P. Liò, and N. Goldman. Molecular phylogenetics: State-of-the-art methods for looking into the past. *Trends Genet.*, 17:262–272, 2001.
51. Z. Yang. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.*, 39:105–111, 1994.
52. Z. Yang. A space-time process model for the evolution of DNA sequences. *Genetics*, 139:993–1005, 1995.
53. J. Yedidia, W. Freeman, and Y. Weiss. Bethe free energy, Kikuchi approximations, and belief propagation algorithms. Technical Report TR2001-16, Mitsubishi Electronic Research Laboratories, 2001.