Variational Upper Bounds for Probabilistic Phylogenetic Models

Ydo Wexler* and Dan Geiger

Dept. of Computer Science, Technion - Israel Institute of Technology, Haifa 32000, Israel
{ywex,dang}@cs.technion.ac.il

Abstract. Probabilistic phylogenetic models which relax the site independence evolution assumption often face the problem of infeasible likelihood computations, for example for the task of selecting suitable parameters for the model. We present a new approximation method, applicable for a wide range of probabilistic models, which guarantees to upper bound the true likelihood of data, and apply it to the problem of probabilistic phylogenetic models. The new method is complementary to known variational methods that lower bound the likelihood, and it uses similar methods to optimize the bounds from above and below. We applied our method to aligned DNA sequences of various lengths from human in the region of the CFTR gene and homologous from eight mammals, and found the upper bounds to be appreciably close to the true likelihood whenever it could be computed. When computing the exact likelihood was not feasible, we demonstrated the proximity of the upper and lower variational bounds, implying a tight approximation of the likelihood.

1 Introduction

Most organisms share a great deal of their genetic code with other forms of life. Phylogenetic tree models are used to associate the genetic makeup of different organisms according to their genetic variation. A node on phylogenetic trees corresponds to a piece of genetic code in a single organism, and the branches and the relative branch lengths measure the relative distance from each organisms' genes to the others. The greater the distance, the more the gene sequence has changed between one organism and the other.

The classical phylogenetic models of Neyman (1971) and Felsenstein (1981) make several assumptions regarding how evolution occurs in the trees, from which the most stringent assumption is that evolution takes place independently in different sites. Over the years more complex probabilistic phylogenetic models have been proposed, which relax the site independence evolution assumption. These complex models that are more biologically realistic, such as the one by Siepel and Haussler (2003), often face the problem of infeasible likelihood computations, for example for the task of selecting suitable parameters for the model. To overcome this problem Jojic et al. (2004) suggested to use variational approximations that lower bound the likelihood of data, and showed that such bounds tend to be close to the true likelihood.

In this paper, we develop tight upper bounds on the likelihood of a given data, that are close to lower bounds so that good estimates of the likelihood become available.

* Corresponding author.

T. Speed and H. Huang (Eds.): RECOMB 2007, LNBI 4453, pp. 226–237, 2007.

Our new approximation method is applicable for a wide range of probabilistic models, including the discussed phylogenetic models. The method assumes a simple distribution Q which approximates the target distribution P of the model, and using Jensen's inequality it upper bounds the likelihood of data with a function of Q and P. The simplicity of Q yields a bound that can be computed efficiently.

Our method is complementary to known variational methods that lower bound the likelihood (e.g. Jordan et al., 1999), and can use an approximating distribution Q suggested by these methods to bound the likelihood also from above.

We applied our method to aligned DNA sequences of various lengths from human in the region of the CFTR gene and homologous from eight mammals, and found the upper bounds to be appreciably close to the true likelihood whenever it could be computed. When computing the exact likelihood was not feasible, we demonstrated the proximity of the upper and lower variational bounds, implying a tight approximation of the likelihood.

The rest of the paper is organized as follows: Section 2 briefly describes phylogenetic HMM models in terms of Bayesian networks or DAG models, and provides a quick overview regarding variational techniques that lower bound the likelihood of data. Section 3 develops our main contribution which are variational upper bounds for probabilistic models such as Bayesian networks. The experimental results are described in Section 4. Finally, we discuss the limitations of variational methods.

2 Preliminaries

We provide background information regarding phylogenetic HMM trees, to which the variational upper bounds suggested herein are applied (Section 2.1), and outline known variational lower bounds of the likelihood of data, which turn out to be close to our upper bounds (Section 2.2).

2.1 Phylogenetic HMM Model

We consider the Phylogenetic HMM model described by Siepel and Haussler (2003). Since the model is given in terms of conditional probabilities, it is convenient to describe it as a DAG model, as done by Jojic et al. (2004). We repeat the description of the model from there with minor changes.

Given a domain of interest having a set of finite variables $\mathbf{s} = (s_1, \ldots, s_n)$ with a positive joint distribution $p(s)$, a DAG model for s is a pair (G, P) where G is a directed acyclic graph and P is a set of conditional probability distributions. A DAG model is also often called a Bayesian network (e.g. Pearl 1988, Jensen 2001). Each node s_i in G corresponds to a variable in s, and to a distribution $p(s_i|\mathbf{pa}(s_i))$, called a local probability distribution, where $\mathbf{pa}(s_i)$ are the parents of s_i in the graph. The joint distribution is given by $p(s) = \prod_{i=1}^{n} p(s_i|\mathbf{pa}(s_i))$. Consequently, the assumed independence relationships between random variables are represented through absence of edges in the model.

A DAG model structure that assumes that evolution takes place independently at each nucleotide site is illustrated in Figure 1a for a simple tree with five species. The

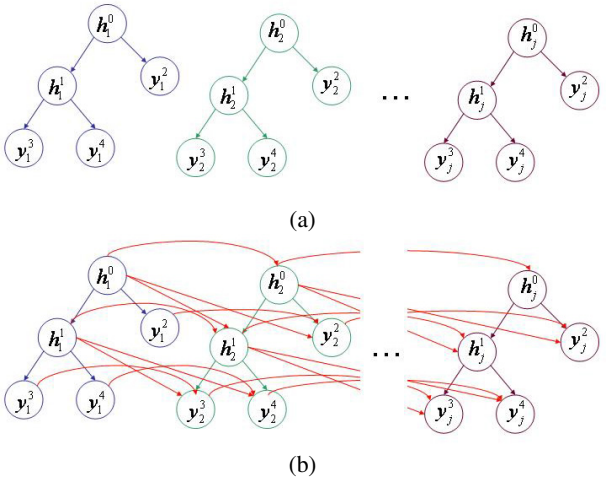

Fig. 1. Probabilistic phylogenetic trees expressed as DAG models. **(a)** The Neyman-Felsenstein tree model that assumes independent evolution in sites. **(b)** The dinucleotide phylogenetic HMM model suggested by Siepel and Haussler (2003).

unknown nucleotide in an ancestor species i at site j is denoted as h^i_j, and the observed nucleotide of an existing species i' at site j' is denoted as $y^{i'}_{j'}$. This is the usual model for which Felsenstein's algorithm for computing likelihood of data is readily applicable. The model of Siepel and Haussler (2003) does not assume that sites are independent, and therefore, edges that connect variables of adjacent sites are added (Figure 1b). This figure illustrates the phylogenetic HMM model of Siepel and Haussler (2003). In this model, a nucleotide of species i at site j depends on the nucleotide of that species at site $j-1$, and its ancestor's nucleotides at sites $j-1$ and j. This model is also called the dinucleotide HMM model, since the two nucleotides of species i and k at site j, where k is the ancestor species of i, are dependent only on the two nucleotides of that species at site $j-1$. Additional more complex models are discussed in Siepel and Haussler (2003).

The local probability distributions of this model are determined by a continuous-time Markov matrix Q of base substitution rates. The matrix Q is of size 16×16, and given evolutionary time t, which is the branch length in the tree, the conditional probabilities $p(s^i_j, s^i_{j-1}|s^k_j, s^k_{j-1})$ are obtained from Q, where k is the ancestor species of i. This distribution then determines the desired probabilities $p(s^i_j|s^i_{j-1}, s^k_j, s^k_{j-1})$. Let $P(t)$ be the matrix of substitution probabilities for branch length t. Then $P(t)$ is given by the solution to the differential equation $\frac{d}{dt}P(t) = P(t)Q$ with initial conditions $P(0) = I$, which is $P(t) = e^{Qt}$. With Q being diagonalizable as $Q = S\Lambda S^{-1}$, the matrix $P(t)$ can be computed as $P(t) = Se^{\Lambda t}S^{-1}$, where $e^{\Lambda t}$ is the diagonal matrix obtained by exponentiating each element on the main diagonal of Λt.

A standard criterion to choose between two DAG models is to prefer a model with higher log-likelihood of the data. However, for the phylogenetic HMM model described here, computing the log-likelihood of data is not feasible, and therefore approximations are needed. In the next section we review known approximations that give lower bounds.

2.2 Variational Lower Bounds

The problem of computing the likelihood, $P(Y = y) = \sum_h P(Y = y, H = h))$, in DAG models is NP-hard (Cooper, 1990; Dagum & Luby, 1993), and although there are many DAG models where exact algorithms are feasible, there are others in which the time and space complexity makes the use of such algorithms infeasible. In these cases fast yet accurate approximations are desired. Herein, we call the task of computing the likelihood by the term inference.

Variational techniques such as the ones suggested by Jordan et al. (1999) are a powerful tool for efficient approximate inference that offers guarantees in the form of lower bounds. In particular, let $P(X)$ be a joint distribution over a set of discrete variables X with the goal to compute the marginal probability $P(Y = y)$, where $Y \subseteq X$. Further assume that this exact computation is not feasible. The idea is to replace P with a distribution Q for which exact inference is feasible, and compute a lower bound for $P(Y = y)$ by using Jensen's inequality:

$$\log P(y) = \log \sum_h Q(h) \frac{P(y,h)}{Q(h)} \geq \sum_h Q(h) \log \frac{P(y,h)}{Q(h)} = -D(Q(H) \,\|\, P(Y{=}y, H))$$

where $H = X \setminus Y$ and $D(\cdot \,\|\, \cdot)$ denotes the KL divergence between two probability distributions.

To obtain tight lower bounds several variational algorithms were devised that try to find an approximating distribution Q which minimizes the KL divergence between Q and the target distribution P ([15,8,17,1,7]). Variational approaches such as the mean field, generalized mean field, and structured mean field differ only with respect to the family of approximating distributions that can be used. Such variational techniques were applied by Jojic et al. (2004) to find lower bounds for the phylogenetic HMM models. The lower bounds computed in the results section herein use a newer algorithm for finding tighter lower bounds suggested by Geiger et al. (2006).

3 Variational Upper Bounds

We denote distributions by $P(x)$ and $Q(x)$, where Q is not necessarily a normalized distribution. Let X be a set of variables and x be an instantiation of these variables. Let $P(x) = \prod_{i=1}^n \Psi_i(d_i)$ and $Q(x) = \prod_{i=1}^n \Phi_i(d_i)$ where d_i is the projection of the instantiation x to the variables in $D_i \subseteq X$, the subsets $\{D_i\}_{i=1}^n$ can overlap, and n is the number of sets D_i. Consider the marginal probability $P(Y = y) = \sum_h P(y, h) = \sum_h \prod_i \Psi_i(d_i)$ where $X = Y \cup H$. We assume throughout that $Q(x)$ is *tractable* in the sense that the marginal probability $Q(Y = y)$ is feasible to compute, while $P(Y = y)$ is not feasible to compute.

We now develop an upper bound for $P(Y = y)$ as summarized in Theorems 1 & 2.

According Jensen's inequality, if f is a concave function and $Z = \{z_1, \ldots, z_n\}$ is a set of real numbers then $f(\sum_{i=1}^n w_i z_i) \geq \sum_{i=1}^n w_i f(z_i)$, where each $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$. By using the concavity of the log function and Jensen's inequality for concave functions, we get the following upper bound:

$$P(Y = y) = \sum_h e^{\log \prod_i \Psi_i(d_i)} = \sum_h e^{\sum_i w_i(h) \log \Psi_i(d_i)^{(1/w_i(h))}} \tag{1}$$

$$\leq \sum_h e^{\log \sum_i w_i(h)\Psi_i(d_i)^{(1/w_i(h))}} = \sum_h \sum_i w_i(h)\Psi_i(d_i)^{(1/w_i(h))}$$

where $\sum_i w_i(h) = 1$ for every instantiation h. Note that this bound can be obtained also by using the weighted power means inequality[1]. Eq. 1 holds with equality regardless of the values of potentials Ψ if and only if

$$w_i(h) = \frac{\log \Psi_i(d_i)}{\log P(h,y)}. \tag{2}$$

Given a tractable distribution $Q(x) = \prod_{i=1}^n \Phi_i(d_i)$ we set $w_i(h) = \frac{\log \Phi_i(d_i)}{\log Q(h,y)}$, which approximates the optimal but intractable choice given by Eq. 2.

With these values for $w_i(h)$, and using the identity $x^{\frac{\log y}{z}} = y^{\frac{\log x}{z}}$, Eq. 1 can be written as:

$$P(Y=y) \leq \sum_h \sum_i \frac{\log \Phi_i(d_i)}{\sum_k \log \Phi_k(d_k)} \prod_m \Phi_m(d_m)^{\frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)}} \tag{3}$$

The upper bound in Eq. 3 holds with equality if Q equals P, because by replacing all occurrences of $\Phi_i(d_i)$ with $\Psi_i(d_i)$ we get

$$P(Y=y) \leq \sum_h \sum_i \frac{\log \Psi_i(d_i)}{\sum_k \log \Psi_k(d_k)} \prod_m \Psi_m(d_m) = \sum_h \prod_m \Psi_m(d_m) = P(Y=y)$$

Eq. 3 remains hard to compute until the sum over h is divided into smaller sums. To obtain a tractable bound we use the arithmetic-geometric means inequality, $\frac{1}{n}\sum_k \log \Phi_k(d_k) \geq \prod_k \log \Phi_k(d_k)^{1/n}$, where $\log \Phi_k(d_k) > 0$. To use this inequality we set all potentials $\Phi_i(d_i)$ to be greater than 1. The resulting tractable upper bound stemming from Eq. 3 is the following:

$$P(Y=y) \leq \frac{1}{n} \sum_h \sum_{i=1}^n \log \Phi_i(d_i) \prod_m \frac{\Phi_m(d_m)^{\frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)}}}{\log \Phi_m(d_m)^{1/n}} \tag{4}$$

Consequently, the following theorem holds.

[1] The weighted power mean $M_w^r(Z)$ of a series of real numbers $Z = \{z_1, \ldots, z_n\}$ is defined for every real $r \in \mathbb{R}$ as

$$M_w^r(z_1, \ldots, z_n) = \begin{cases} \left[\sum_{i=1}^n w_i z_i^r\right]^{1/r} & \text{if } r \neq 0 \\ \prod_{i=1}^n z_i^{w_i} & \text{if } r = 0 \end{cases}$$

where $w_1, \ldots, w_n$ are positive real numbers such that $\sum_{i=1}^n w_i = 1$. Note that $M_w^r(Z) \xrightarrow{r \to 0} M_w^0(Z)$.

The power mean inequality states that for two real numbers s, t, the relation $s < t$ implies $M_w^s < M_w^t$, and the upper bounds are obtained by setting $s = 0$, $t = 1$, and $z_i = \Psi_i(d_i)^{(1/w_i)}$.

Theorem 1 (upper bound). *Let H and Y be two disjoint sets of variables such that $H \cup Y = X$, and let $P(x)$ and $Q(x)$ be distributions that factor according to $P(x) = \prod_{i=1}^{n} \Psi_i(d_i)$ and $Q(x) = \prod_{i=1}^{n} \Phi_i(d_i)$ where d_i is the projection of the instantiation x to the variables in $D_i \subseteq X$. Then the following is an upper bound on $P(Y = y)$,*

$$P(Y=y) \leq \frac{1}{n} \sum_{i} \sum_{D_i} \log \Phi_i(d_i) \left[\sum_{h \setminus D_i} \prod_{m} \frac{\Phi_m(d_m)^{\frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)}}}{\log \Phi_m(d_m)^{1/n}} \right] \tag{5}$$

Proof: The proof is immediate from Eq. 4 where we replace the sums over i and h, and divide the sum over h such that first we sum over variables in D_i and then over the rest of the variables in H. □

Assuming that $M = \max_i\{|D_i|\}$ is at most a given constant, the time needed to compute the bound given in Eq. 5 is linear in the number of variables in the model and proportional to the time needed to compute $Q(y)$. Therefore, the tractability of this bound is a direct consequence of the assumption of tractable inference on distribution Q.

Since the maximal size M of the sets in the model can sometime be large enough to significantly slow computations of the upper bound, we develop a more efficient method to compute the upper bound that does not depend on M. To do so, we use the following lemma.

Lemma 1. *Given two sets of positive real numbers $X = \{x_1 \ldots, x_n\}$ and $Y = \{y_1 \ldots, y_n\}$ and a positive real number r, the following inequalities hold.*
If $0 < r \leq 1$, then

$$\sum_{i=1}^{n} \frac{x_i^r}{y_i} \leq \left(\sum_{i=1}^{n} \frac{x_i}{y_i} \right)^{r} \cdot \left(\sum_{i=1}^{n} y_i^{-1} \right)^{1-r}.$$

If $1 \leq r < 2$, then

$$\sum_{i=1}^{n} \frac{x_i^r}{y_i} \leq \left(\sum_{i=1}^{n} \frac{x_i}{y_i} \right)^{2-r} \cdot \left(\sum_{i=1}^{n} \frac{x_i^2}{y_i} \right)^{r-1}.$$

For $r = 1$ equalities hold.

Proof: We use the Euclidean case of Hölder's inequality, stating that for two sets of positive real numbers $X = \{x_1 \ldots, x_n\}$ and $Y = \{y_1 \ldots, y_n\}$, and for two real numbers $p, q \geq 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$\sum_{i=1}^{n} x_i \cdot y_i \leq \left(\sum_{i=1}^{n} x_i^p \right)^{1/p} \cdot \left(\sum_{j=1}^{n} y_j^q \right)^{1/q}.$$

For $0 < r \leq 1$, we get using Hölder's inequality,

$$\sum_{i=1}^{n} \frac{x_i^r}{y_i} = \sum_{i=1}^{n} \left(\frac{x_i}{y_i} \right)^{r} \cdot y_i^{r-1} \leq \left(\sum_{i=1}^{n} \left(\frac{x_i}{y_i} \right)^{r \cdot p} \right)^{1/p} \cdot \left(\sum_{i=1}^{n} y_i^{(r-1) \cdot q} \right)^{1/q}.$$

Setting $p = \frac{1}{r}$ and $q = \frac{1}{1-r}$ we get

$$\sum_{i=1}^{n} \frac{x_i^r}{y_i} \leq \left(\sum_{i=1}^{n} \frac{x_i}{y_i} \right)^r \cdot \left(\sum_{i=1}^{n} y_i^{-1} \right)^{1-r}.$$

Similarly, for $1 \leq r < 2$, we get using Hölder's inequality,

$$\sum_{i=1}^{n} \frac{x_i^r}{y_i} = \sum_{i=1}^{n} \left(\frac{x_i}{y_i} \right)^{2-r} \cdot \left(\frac{x_i^2}{y_i} \right)^{r-1} \leq \left(\sum_{i=1}^{n} \left(\frac{x_i}{y_i} \right)^{(2-r)\cdot p} \right)^{1/p} \cdot \left(\sum_{i=1}^{n} \left(\frac{x_i^2}{y_i} \right)^{(r-1)\cdot q} \right)^{1/q}.$$

Setting $p = \frac{1}{2-r}$ and $q = \frac{1}{r-1}$ we get

$$\sum_{i=1}^{n} \frac{x_i^r}{y_i} \leq \left(\sum_{i=1}^{n} \frac{x_i}{y_i} \right)^{2-r} \cdot \left(\sum_{i=1}^{n} \frac{x_i^2}{y_i} \right)^{r-1}.$$

□

Theorem 2 (Efficient upper bound). *Let H and Y be two disjoint sets of variables such that $H \cup Y = X$, and let $P(x)$ and $Q(x)$ be distributions that factor according to $P(x) = \prod_{i=1}^{n} \Psi_i(d_i)$ and $Q(x) = \prod_{i=1}^{n} \Phi_i(d_i)$ where $\Psi_i > 1$, $\Phi_i > 1$ and $\frac{\log \Psi_i}{\log \Phi_i} < 2$ for every $i = 1, \ldots, n$, and where d_i is the projection of the instantiation x to the variables in $D_i \subseteq X$. In addition, let U_i denote the set of instantiations of D_i for which $\Phi_i(d_i) \leq \Psi_i(d_i)$, and let L_i denote the rest of instantiations of D_i. Then the following is an upper bound on $P(Y = y)$,*

$$P(Y = y) \leq \frac{1}{n} \sum_i \left[\sum_{d_i \in L_i} \log \Phi_i(d_i) \Lambda_{L_i} + \sum_{d_i \in U_i} \log \Phi_i(d_i) \Lambda_{U_i} \right] \tag{6}$$

where

$$\Lambda_{L_i} = \left(\sum_{h \setminus D_i} \prod_m \frac{\Phi_m(d_m)}{\log \Phi_m(d_m)^{1/n}} \right)^{\frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)}} \cdot \left(\sum_{h \setminus D_i} \prod_m \frac{1}{\log \Phi_m(d_m)^{1/n}} \right)^{1 - \frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)}}$$

and

$$\Lambda_{U_i} = \left(\sum_{h \setminus D_i} \prod_m \frac{\Phi_m(d_m)}{\log \Phi_m(d_m)^{1/n}} \right)^{2 - \frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)}} \cdot \left(\sum_{h \setminus D_i} \prod_m \frac{\Phi_m(d_m)^2}{\log \Phi_m(d_m)^{1/n}} \right)^{\frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)} - 1}$$

Proof: Lemma 1 implies that when $\Phi_i(d_i) \geq \Psi_i(d_i) > 1$, we can replace every bracketed term $\sum_{h \setminus D_i} \prod_m \left[\Phi_m(d_m)^{\frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)}} / \log \Phi_m(d_m)^{1/n} \right]$ in Eq. 5 with Λ_{L_i} and when $1 < \Phi_i(d_i) < \Psi_i(d_i)$, we can replace it with Λ_{U_i}, since $\frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)} < 2$. □

Computing each term, Λ_{U_i} or Λ_{L_i}, involves only two sums of products, where each sum factors according to distribution Q. These computations can be performed by using any

algorithm such as *bucket elimination algorithm* or the *sum-product algorithm* described by Dechter (1999) and Kschischang, Frey & Loeliger (2001) . According to Eq. 6 only a linear number of calls to such procedures are needed to obtain the upper bound.

If each potential Ψ_i and Φ_i is multiplied by a large factor α, all the terms $\frac{\log \Psi_i}{\log \Phi_i}$ approach one as α grows. This reduces the accuracy gap when using Hölder's inequality in Eq. 6 with $r = \frac{\log \Psi_i}{\log \Phi_i}$. In addition, note that multiplying the potentials Φ_i by α also serves the tightness of the arithmetic-geometric inequality used to obtain Eq. 5, since for each pair of potentials Φ_j and Φ_k, the ratio $\frac{\log \Phi_j}{\log \Phi_k}$ approaches one as α grows. A large enough α guarantees that $\frac{\log \Psi_i}{\log \Phi_i} < 2$ for all sets D_i and thus the applicability of Theorem 2. In our experiments we use $\ln \alpha = 300$.

4 Approximations for Phylogenetic HMM Models

The dinucleotide phylogenetic HMM model of Siepel and Haussler (2003), described in Section 2.1, lead to improvements over previous models in several biological tasks such as gene finding. But, despite its enhanced power, it also requires evaluating an intractable likelihood for the purpose of finding optimal parameters for the model. Jojic et al. (2004) used variational techniques, similar to the ones described in Section 2.2 to lower bound the likelihood of data, and showed that when the exact likelihood can be computed (although with much effort), the approximations were tight.

We use the upper bounds suggested in Section 3 to compute the likelihood of phylogenetic trees with a small error, by bounding it tightly from above and below. First, we show the upper bounds are close to the true likelihood when this can be computed. Then, for larger phylogenetic trees, where computing the exact likelihood is infeasible, we show the proximity of the lower and upper bounds. To set a tractable approximating distribution Q, we use a parameter k which determines its topology: sets that contain variables from sites ck and $ck+1$, for $c = 1, 2, 3, \ldots$, are split into two disjoint subsets, D_{i1} and D_{i2}, where D_{i1} contains only variables in D_i from site ck and D_{i2} contains the rest of the variables in D_i. Their respective potentials $\Phi_i(d_i)$ therefore factor according to $\Phi_i(d_i) = \Phi_{i1}(d_{i1})\Phi_{i2}(d_{i2})$. In our experiments we used $k = 10$ when computing the exact likelihood was feasible and $k = 5$ when the likelihood computation was infeasible. The lower bounds were obtained by using a recent variational algorithm called VIP* (Geiger et al., 2006).

We repeat each upper bound computation twice, with the difference of the way potentials Φ_i are chosen. The first choice is what we call non-informative (NI), where each potential $\Phi_i(d_i) = \prod_{j=1}^{m_i} \Phi_{ij}(d_{ij})$ is a product of m_i sub-potentials of sets $D_{ij} \subseteq D_i$. A sub-potential $\Phi_{ij}(d_{ij})$ is set to be the $1/m_i$ power of the average value of $\Psi_i(d_i)$ of all instantiations d_i consistent with d_{ij}. More formally, $\Phi_{ij}(d_{ij}) = \left(\frac{1}{|C_{d_{ij}}|} \sum_{d_i \in C_{d_{ij}}} \Psi(d_i)\right)^{1/m_i}$ where $C_{d_{ij}}$ is the set of instantiations d_i consistent with d_{ij}.

The second choice of potentials, called variational-based (VB), is based on variational algorithms, such as VIP*, that optimize the approximating distribution Q in order

to set tight lower bounds on the likelihood. If the topology of Q given for these algorithms follows the factorization suggested in Section 3 (i.e. every potential Ψ_i in P has its corresponding potential Φ_i in Q), the potentials found by these optimization algorithms to lower bound the likelihood can also serve to upper bound it using the method proposed herein.

We ran the tests on data used by Siepel and Haussler (2003) that contains sequences from human in the region of the CFTR gene and homologous from eight mammals: chimp, baboon, cow, pig, cat, dog, mouse and rat. The sequences are aligned, and we used portions of this alignment to obtain our results. The substitution probabilities in all models were computed from the dinucleotide substitution matrix obtained by Jojic et al. (2004), and the branch lengths in each tree were randomly chosen, normally distributed around predetermined means. The first tests used two data sets, similar to those used by Jojic et al. (2004), where each set consisted of three sequences. The sequences in set A were taken from the cow, mouse and human genomes and were of length 30Knc, and the sequences in set B were taken from the cow, pig and dog genomes and were of length 20Knc. Figure 2a and 2b plot the upper bounds versus the exact log-likelihoods of trees with different branch lengths. Lower bounds are also shown in the figure to demonstrate the tightness level of these bounds. The average differences for the trees in set A between the upper bounds and the exact likelihoods were 1% for the NI method

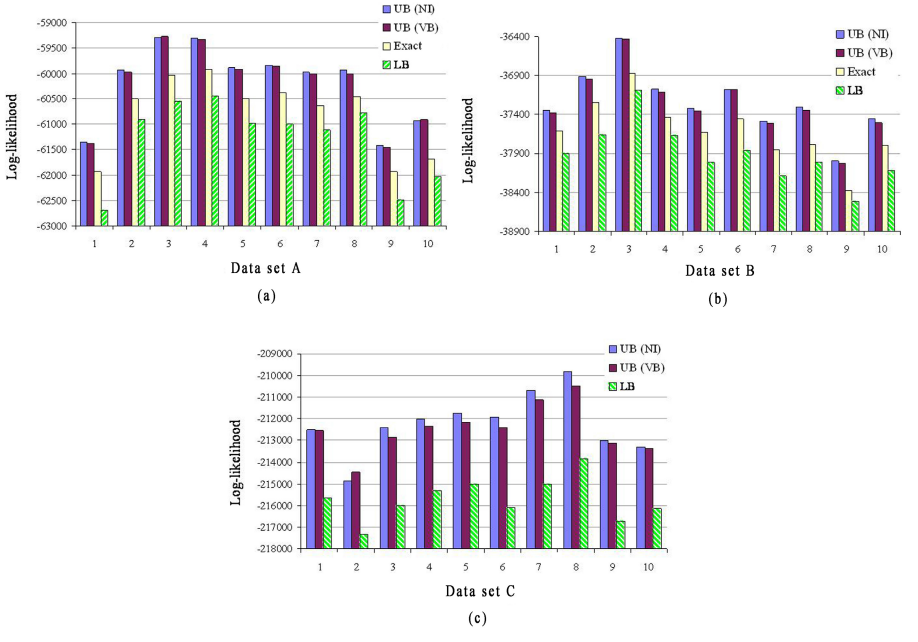

Fig. 2. Upper and lower bounds on the likelihood of data of phylogenetic HMM models for sets A, B and C with different branch lengths. **(a) & (b)** Bounds versus the exact likelihood for models of sets A and B. **(c)** Bounds for models of set C, for which computing the exact likelihood is infeasible.

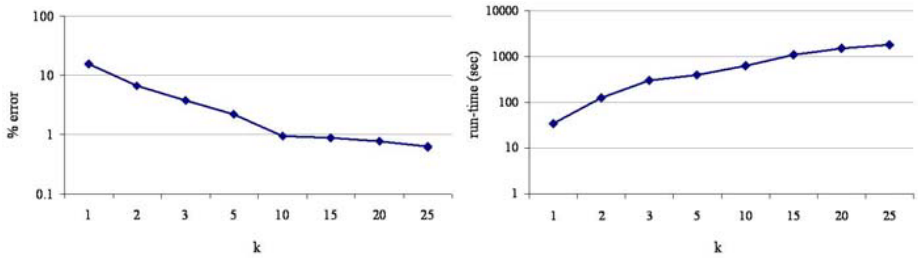

Fig. 3. Accuracy and run-time as a function of parameter k of decomposing the model. **(a)** Accuracy as a function of k. **(b)** Run-time as a function of k.

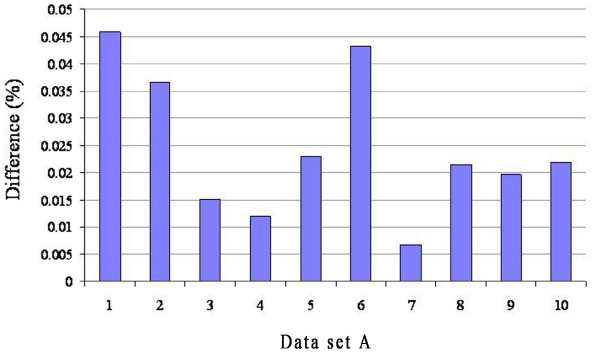

Fig. 4. The difference in accuracy between upper bounds computed via Eq. 5 and bounds computed via Eq. 6

and 0.95% for the VB method, and for trees in set B the average differences were 0.97% (NI) and 0.9% (VB).

The upper and lower bounds for an additional set of aligned sequences that contained sequences of length 30Knc from all nine organisms (Set C) are illustrated in Figure 2c. For this set it is infeasible to compute the exact likelihood, but the proximity of the upper and lower bounds allows us to predict the likelihood with a small error. The NI method yielded an average of 1.64% difference from the lower bounds and the VB method yielded an average of 1.52% from the lower bounds for the models in this set.

As shown in Figure 2, both choices of potentials (NI and VB) performed similarly, with a small advantage of the VB method over NI in most experiments. In other experiments we performed, we found that arbitrary choice of potentials often lead to significant decrease in the tightness of the bounds (up to 45%), and therefore an algorithm is desired to find potentials that lead to tight bounds.

The parameter k used for decomposing the tree model into parts of k sites is a trade-off between run-time and accuracy: the larger k is the more time consuming it is to compute the upper bounds, however, the bounds computed are also more accurate. The

default value of k was set to 10 for trees in Set A. Figure 3 shows the results for these trees as a function of k in terms of accuracy and in terms of run-time.

Finally, we tested the difference in accuracy between upper bounds computed via Eq. 5 and those computed via Eq. 6. The expected run-time ratio between these two methods is the average probability table size in the model. Since no preprocessing such as summing over some variables was executed, the expected ratio was 81.25. As shown in Figure 4, the differences in accuracy of the upper bounds were negligible, less than 0.05% of their log value, when applied to phylogenetic trees in data set A. This implies that when the size of the probability tables is large, Eq. 6 is an attractive and efficient alternative to Eq. 5.

5 Discussion

Computing the likelihood of many probabilistic models is infeasible and calls for efficient approximations. Our results on phylogenetic models show that the suggested upper bounds are appreciably tight and together with other variational methods allow to compute the likelihood almost exactly in feasible time. We have also started using the upper bounds to approximate other probabilistic models and believe that they can be applied to a wide range of models and for various tasks. One additional task we explore is bounding the MAP assignment probability in order to set optimal parameters for models where finding the exact MAP assignment is infeasible. The goodness of the bounds heavily depends on the choice of an approximating distribution Q, and more work on choosing useful Q functions is desired, as indicated by Xing et al. (2004).

As with variational methods that offer lower bounds on the likelihood, if the dependence of variables under Q largely differs from their dependence under the target distribution P, these methods yield loose bounds. When exploring probabilistic models to genetic linkage analysis, as used by Fishlson and Geiger (2002), we found that the variational methods we used did not offer sufficiently good approximating distributions for these models, and therefore did not give tight enough bounds. Geiger et al. (2006) provided results of variational techniques on genetic linkage analysis problems and showed that although the lower bounds followed the shape of the likelihood function, the difference from the true log-likelihood reached 20%. The difficulty in finding good approximations to this model may lie in the level of determinism of the model: relaxing deterministic dependence relationships between variables reduced accuracy far more than when relaxing mild dependence relationships. When computing the upper bounds suggested herein for genetic linkage analysis, the results were within 10% from the true log-likelihood.

Ackowledgements

The research is supported by the Israel Science Foundation and the Israeli Science Ministry.

References

1. C. Bishop and J. Winn. Structured variational distributions in VIBES. In *Artificial Intelligence and Statistics*, Key West, Florida, USA, 2003. Society for Artificial Intelligence and Statistics.
2. G. Cooper. Probabilistic inference using belief networks is NP-hard. *Artificial Intelligence*, 42:393–405, 1990.
3. P. Dagum and M. Luby. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60(1):141–153, 1993.
4. R. Dechter. Bucket elimination: A unifying framework for reasoning. *Artificial Intelligence*, 113(1-2):41–85, 1999.
5. J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.
6. M. Fishelson and D. Geiger. Exact genetic linkage computations for general pedigrees. *Bioinformatics*, 18:S189–S198, 2002.
7. D. Geiger, C. Meek, and Y. Wexler. A variational inference procedure allowing internal structure for overlapping clusters and deterministic constraints. *Journal of Artificial Intelligence Research (JAIR)*, 27:1–23, 2006.
8. Z. Ghahramani and M. I. Jordan. Factorial hidden Markov models. *Machine Learning*, 29:245–273, 1997.
9. F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer-Verlag New York, Inc., 2001.
10. V. Jojic, N. Jojic, C. Meek, D. Geiger, A. Siepel, D. Haussler, and D. Heckerman. Efficient approximations for learning phylogenetic HMM models from data. *Bioinformatics*, 20:161–168, 2004.
11. M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and S. L. K. *An introduction to variational methods for graphical models*. Learning Graphical Models. MIT Press, 1999.
12. F. R. Kschischang, B. J. Frey, and H. A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on information theory*, 47(2):498–519, February 2001.
13. J. Neyman. Molecular studies of evolution: a source of novel statistical problems. In *S. S. Gupta and J. Yackel (eds), Statistical desicion theory and related topics*, pages 1–27. Academic Press, New York, 1971.
14. J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, San Mateo, CA, 1988.
15. L. K. Saul and M. I. Jordan. Exploiting tractable substructures in intractable networks. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 8, pages 486–492. The MIT Press, 1996.
16. A. Siepel and D. Haussler. Combining phylogenetic and hidden markov models in biosequence analysis. In *RECOMB '03: Proceedings of the seventh annual international conference on Research in computational molecular biology*, pages 277–286, New York, NY, USA, 2003. ACM Press.
17. W. Wiegerinck. Variational approximations between mean field theory and the junction tree algorithm. In *UAI '00: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 626–633, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
18. E. P. Xing, M. I. Jordan, and S. Russell. Graph partition strategies for generalized mean field inference. In *AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 602–610, Arlington, Virginia, United States, 2004. AUAI Press.