JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 15, Number 7, 2008

Pp. 721–735
DOI: 10.1089/cmb.2007.0129

Variational Upper and Lower Bounds for Probabilistic Graphical Models

YDO WEXLER and DAN GEIGER

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 and lower 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 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.

Key words: algorithms, computational molecular biology, genetic mapping, learning, secondary structure.

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 in 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 organism's 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 at 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

Computer Science Department, Technion, Haifa, Israel.

infeasible likelihood computations, for example, for the task of selecting suitable parameters for the model. To overcome this problem, Jojic et al. (2004) suggested we 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 and lower bounds on the likelihood of a given data, so that good estimates of the likelihood become available. 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 the weighted power-mean inequality it upper bounds the likelihood of data with a function of Q and P. Combining the resulting bounds with an inequality derived by Pečarić and Mićić (2005), new lower bounds also become available. The simplicity of Q yields bounds that can be computed efficiently.

Our method is complementary to known variational methods that lower bound the likelihood (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. We also developed similar upper bounds for computing the MPE probability and applied them to medical image reconstruction.

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 bound the likelihood of data from below. Section 3 develops our main contribution which are variational upper and lower bounds for probabilistic models such as Bayesian networks. The experimental results are described in Section 4. Finally, we discuss the limitations of variational methods. The bounds on the MPE probability and their application to medical images is described in the appendix.

2. PRELIMINARIES

We provide background information regarding phylogenetic HMM trees, to which the variational 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 (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 unknown nucleotide in an ancestor species i at site j is denoted as h_j^i, and the observed nucleotide of an existing species i' at site j' is denoted as $y_{j'}^{i'}$. 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 (Fig. 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.

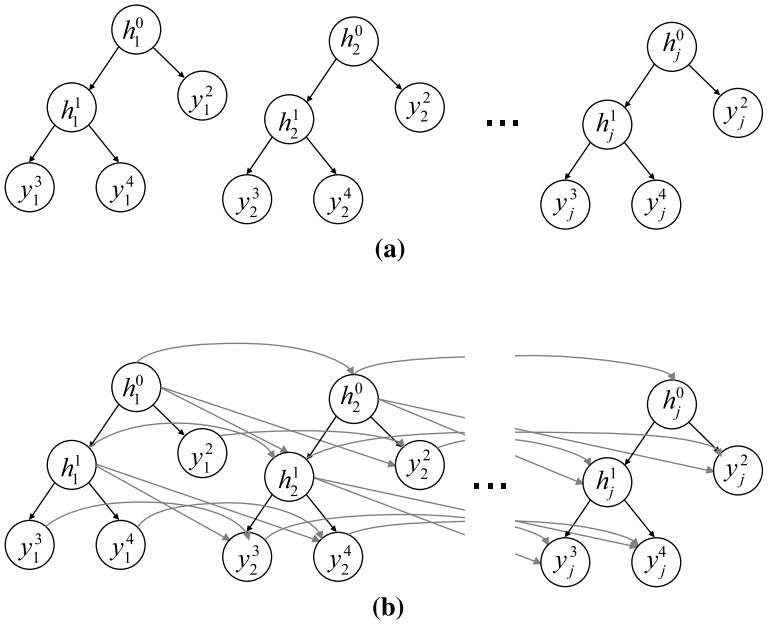

FIG. 1. Probabilistic phylogenetic trees expressed as DAG models. **(a)** The Neyman-Felsenstein tree model that assumes independent evolution in sites. **(b)** The dinucleotide phylogenetic HMM model suggested by Siepel and Haussler (2003).

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 B of base substitution rates. The matrix B is of size 16×16, and given evolutionary time t, which is the branch length in the tree, the conditional probabilities $p(s_j^i, s_{j-1}^i | s_j^k, s_{j-1}^k)$ are obtained from B, where k is the ancestor species of i. This distribution then determines the desired probabilities $p(s_j^i | s_{j-1}^i, s_j^k, s_{j-1}^k)$. 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 $\frac{d}{dt}P(t) = P(t)B$ with initial conditions $P(0) = I$, which is $P(t) = e^{Bt}$. With B being diagonalizable as $B = S\Lambda S^{-1}$, 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 and 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 that minimizes the KL divergence between Q and the target distribution P (Saul and Jordan, 1996; Ghahramani and Jordan, 1997; Wiegerinck, 2000; Bishop and Winn, 2003; Geiger et al., 2006). 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. This type of lower bounds computed in the results section herein use a newer algorithm for finding tighter lower bounds suggested by Geiger et al. (2006).

3. NEW VARIATIONAL UPPER AND LOWER 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.

3.1. Upper bounds

We now develop upper bounds for $P(Y = y)$ as summarized in Theorems 1 and 2. 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 the weights $w_1, \ldots, w_n$ are positive real numbers such that $\sum_{i=1}^n w_i = 1$. Note that $M_w^r(Z) \overset{r \to 0}{\longrightarrow} M_w^0(Z)$.

Let s and t be two real numbers such that $s < t$, then according to the power mean inequality $M_w^s \leq M_w^t$. Using the power mean inequality with $s = 0$, $t = 1$ and $z_i = \Psi_i(d_i)^{(1/w_i)}$, we obtain the following upper bound:

$$P(Y = y) = \sum_h \prod_i \left[(\Psi_i(d_i))^{(1/w_i(h))} \right]^{w_i(h)} \leq \sum_h \sum_i w_i(h) \, (\Psi_i(d_i))^{(1/w_i(h))} \tag{1}$$

where $\sum_i w_i(h) = 1$ for every instantiation h. Note that this bound can be obtained also via Jensen's inequality stating that 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, bounds identical to those in Eq. 1 are obtained via:

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

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

Eq. 1 holds with equality regardless of the values of potentials Ψ_i if and only if

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

However, this optimal choice leads to an intractable upper bounds in Eq. 5. Instead, 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.

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 obtain 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 obtain

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

Similarly, for $1 \leq r < 2$, we obtain 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 obtain

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

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 normalized distributions with corresponding unnormalized distributions* $\tilde{P}(X)$ *and* $\tilde{Q}(X)$ *that factor according to* $P(x) = \frac{1}{K}\tilde{P}(X) = \prod_{i=1}^{n} \tilde{\Psi}_i(d_i)$ *and* $Q(x) = \frac{1}{K}\tilde{Q}(X) = \prod_{i=1}^{n} \tilde{\Phi}_i(d_i)$ *where* $\tilde{\Psi}_i > 1$*,* $\tilde{\Phi}_i > 1$ *and* $\frac{\log \tilde{\Psi}_i}{\log \tilde{\Phi}_i} < 2$ *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* $\tilde{\Phi}_i(d_i) \leq \tilde{\Psi}_i(d_i)$*, and let* L_i *denote the rest of instantiations of* D_i*. Then the following is an upper bound on* $P(Y = y)$*,*

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

where

$$\Lambda_{L_i} = \left(\sum_{h \setminus D_i} \prod_m \frac{\tilde{\Phi}_m(d_m)}{\left[\log \tilde{\Phi}_m(d_m) \right]^{1/n}} \right)^{\frac{\log \tilde{\Psi}_i(d_i)}{\log \tilde{\Phi}_i(d_i)}} \cdot \left(\sum_{h \setminus D_i} \prod_m \frac{1}{\left[\log \tilde{\Phi}_m(d_m) \right]^{1/n}} \right)^{1 - \frac{\log \tilde{\Psi}_i(d_i)}{\log \tilde{\Phi}_i(d_i)}}$$

and

$$\Lambda_{U_i} = \left(\sum_{h \setminus D_i} \prod_m \frac{\tilde{\Phi}_m(d_m)}{\left[\log \tilde{\Phi}_m(d_m)\right]^{1/n}} \right)^{2 - \frac{\log \tilde{\Psi}_i(d_i)}{\log \tilde{\Phi}_i(d_i)}} \cdot \left(\sum_{h \setminus D_i} \prod_m \frac{\tilde{\Phi}_m(d_m)^2}{\left[\log \tilde{\Phi}_m(d_m)\right]^{1/n}} \right)^{\frac{\log \tilde{\Psi}_i(d_i)}{\log \tilde{\Phi}_i(d_i)} - 1}$$

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

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 et al. (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 to obtain the potentials $\tilde{\Psi}_i$ and $\tilde{\Phi}_i$ respectively, all the terms $\frac{\log \tilde{\Psi}_i}{\log \tilde{\Phi}_i}$ approach one as α grows. This reduces the accuracy gap when using Hölder's inequality in Eq. 6 with $r = \frac{\log \tilde{\Psi}_i}{\log \tilde{\Phi}_i}$. In addition, note that multiplying the potentials Φ_i by α also improves the tightness of the arithmetic-geometric inequality used to obtain Eq. 5, since for each pair of potentials $\tilde{\Phi}_j$ and $\tilde{\Phi}_k$, the ratio $\frac{\log \tilde{\Phi}_j}{\log \tilde{\Phi}_k}$ approaches one as α grows. A large enough α guarantees that $\frac{\log \tilde{\Psi}_i}{\log \tilde{\Phi}_i} < 2$ for all sets D_i and thus the applicability of Theorem 2. In our experiments, we use $\ln \alpha = 300$.

We applied the weighted power means inequality for a specific parameter $r = 1$. Choosing a smaller parameter $r = 1/l$ where l is an integer greater than 1 yields a tighter bound but at the cost of additional run-time. The general form of the bounds, which replaces Eq. 1, is as follows:

$$P(Y = y) = \sum_h \prod_i \Psi_i(d_i) \leq \sum_h \left[\sum_i w_i \Psi_i(d_i)^{\frac{1}{l \cdot w_i(d_i)}} \right]^l = \sum_h \sum_{i_1 \ldots i_l} \prod_{t=1}^{l} w_{i_t}(d_{i_t}) \Psi_{i_t}(d_{i_t})^{\frac{1}{l \cdot w_{i_t}(d_{i_t})}}$$

Weights that are chosen according to Eq. 2 are still optimal for the general case. Choosing the weights as before, and applying the arithmetic-geometric inequality, the bounds can be further rewritten:

$$P(Y = y) \leq \sum_{i_1 \ldots i_l} \sum_{D_{i_1} \ldots D_{i_l}} \left[\prod_{t=1}^{l} \log \Phi_{i_t}(d_{i_t}) \right] \sum_{h \setminus \{D_{i_1} \ldots D_{i_l}\}} \frac{1}{\left[\sum_k \log \Phi_k(d_k)\right]^l} \prod_m \Phi_m(d_m)^{\frac{\log \prod_{t=1}^{l} \Psi_{i_t}(d_{i_t})}{l \cdot \log \prod_{t=1}^{l} \Phi_{i_t}(d_{i_t})}} \tag{7}$$

$$\leq \frac{1}{n^l} \sum_{i_1 \ldots i_l} \sum_{D_{i_1} \ldots D_{i_l}} \left[\prod_{t=1}^{l} \log \Phi_{i_t}(d_{i_t}) \right] \sum_{h \setminus \{D_{i_1} \ldots D_{i_l}\}} \prod_m \frac{\Phi_m(d_m)^{\frac{\log \prod_{t=1}^{l} \Psi_{i_t}(d_{i_t})}{l \cdot \log \prod_{t=1}^{l} \Phi_{i_t}(d_{i_t})}}}{\left[\log \Phi_m(d_m)\right]^{l/n}}$$

As can be seen from the equations, the run-time is exponential in l.

3.2. Lower bounds

The method described in Section 3.1 for obtaining upper bounds can also be used, with small modifications, to obtain lower bounds on the likelihood. In particular, we use an inequality which was originally described by Pečarić and Mićić (2005): Let $s, t \in \mathbb{R}$ be two non-zero numbers such that $s \leq t$ and $t \geq 1$ and $-1 < s \leq 1$, and $k = \frac{M}{m}$ where M and m are the maximum and minimum numbers in a series of real numbers A, respectively. Then,

$$\Delta(k, s, t)^{-1} M_w^t(A) \leq M_w^s(A) \tag{8}$$

where $M_w^r(A)$ is the r^{th} weighted power mean of the series A and $\Delta(k,s,t) = \left(\frac{s(k^t-k^s)}{(t-s)(k^s-1)}\right)^{1/t} \left(\frac{t(k^s-k^t)}{(s-t(k^t-1)}\right)^{-1/s}$.

Let Q be a tractable distribution $Q(x) = \prod_{i=1}^n \Phi_i(d_i)$ as before, with respective weights $w_i(h) = \frac{\log \Phi_i(d_i)}{\log Q(h,y)}$, and let $Z(h) = \{z_1(h), \ldots, z_n(h)\}$ such that $z_i(h) = \Psi_i(d_i)^{1/w_i(h)}$, then for a small positive number ϵ we get:

$$P(Y = y) = \sum_h M_w^0(Z(h)) \geq \sum_h M_w^{-\epsilon}(Z(h)) \geq \sum_h \Delta(k(h), -\epsilon, 1)^{-1} M_w^1(Z(h)) \tag{9}$$

where the last inequality is due to Eq. 8.

For a tractable bound, we set two virtual potentials $\Psi_m = (1+\epsilon_m)$ and $\Psi_M = \epsilon_M > 1$ and their respective weights w_m and w_M, such that every term $z_i(h)$ is in the range $z_m \leq z_i(h) \leq z_M$ for every instantiation h where $z_m = \Psi_m^{(1/w_m)}$ and $z_M = \Psi_M^{(1/w_M)}$. The condition $z_m \leq z_i(h) \leq z_M$ is attainable for every set of potentials by multiplying the original potentials Ψ_i by a constant and setting ϵ_m and ϵ_M and their weights accordingly. Now we can rewrite Eq. 9 as follows:

$$P(Y = y) = \sum_h M_w^0(Z(h)) \geq \sum_h M_w^{-\epsilon}(Z(h)) \geq \sum_h \Delta\left(\frac{z_M}{z_m}, -\epsilon, 1\right)^{-1} M_w^1(Z(h)) \tag{10}$$

Since $\Delta(\frac{z_M}{z_m}, -\epsilon, 1)^{-1}$ does not depend on the instance h, this inequality can be written as:

$$P(Y = y) \geq \Delta(\frac{z_M}{z_m}, -\epsilon, 1)^{-1} \sum_h M_w^1(Z(h)) \tag{11}$$

$$= \Delta\left(\frac{z_M}{z_m}, -\epsilon, 1\right)^{-1} \sum_h \sum_i \frac{\log \Phi_i(d_i)}{\sum_k \log \Phi_k(d_k)} \prod_j \Phi_j(d_j)^{\frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)}}. \tag{12}$$

We denote the maximal potential in Q by $\Phi_{max} = \max_{h,k}\{\Phi_k(d_k)\}$. Hence, since $\sum_k \log \Phi_k(d_k) \leq n \log \Phi_{max}$, we can lower bound the expression in Eq. 11 to obtain

$$P(Y = y) \geq \frac{\Delta\left(\frac{z_M}{z_m}, -\epsilon, 1\right)^{-1}}{n \log \Phi_{max}} \sum_i \sum_{D_i} \log \Phi_i(d_i) \left[\sum_{h \setminus D_i} \prod_j \Phi_j(d_j)^{\frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)}}\right]. \tag{13}$$

We note that, if each potential Ψ_i and Φ_i is multiplied by a large factor α, as done before, the decrease in the tightness of the bound as a result of using $\log \Phi_{max}$ reduces as α grows. In addition, we note that these proposed bounds show results comparable with known variational lower bounds on the models we tried, and when computing both, one can take the maximum of the two to obtain tighter bounds.

4. APPROXIMATIONS FOR PHYLOGENETIC HMM MODELS

The dinucleotide phylogenetic HMM model of Siepel and Haussler (2003), described in Section 2.1, leads 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 and lower 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 demonstrate a small gap between the lower and upper bounds. To set a tractable approximating distribution Q, we use a parameter k that determines its structure: 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 $\Phi_i(d_i)$ then 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 both by using a recent variational algorithm called VIP* (Geiger et al., 2006), and by using the lower bounds suggested in Section 3.2.

We computed upper bounds using two choices of potentials Φ_i. 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 $\Phi_{ij}(d_{ij}) = \left(\frac{1}{|C_{d_{ij}}|} \sum_{d_i \in C_{d_{ij}}} \Psi(d_i)\right)^{1/m_i}$, 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 3.1 (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. This choice of potentials is also used when computing the lower bounds proposed in Section 3.2.

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 30K bp (namely, each sequence consists of 30,000 symbols which are either A, C, G, or T), and the sequences in set B were taken from the cow, pig, and dog genomes and were of length 20K bp. Figure 2a,b plots the upper and lower bounds versus the exact log-likelihoods of trees with different branch lengths.

The upper and lower bounds for an additional set of aligned sequences that contained sequences of length 30K bp from all nine organisms (Set C) are illustrated in Figure 2c. For this set, it is infeasible to

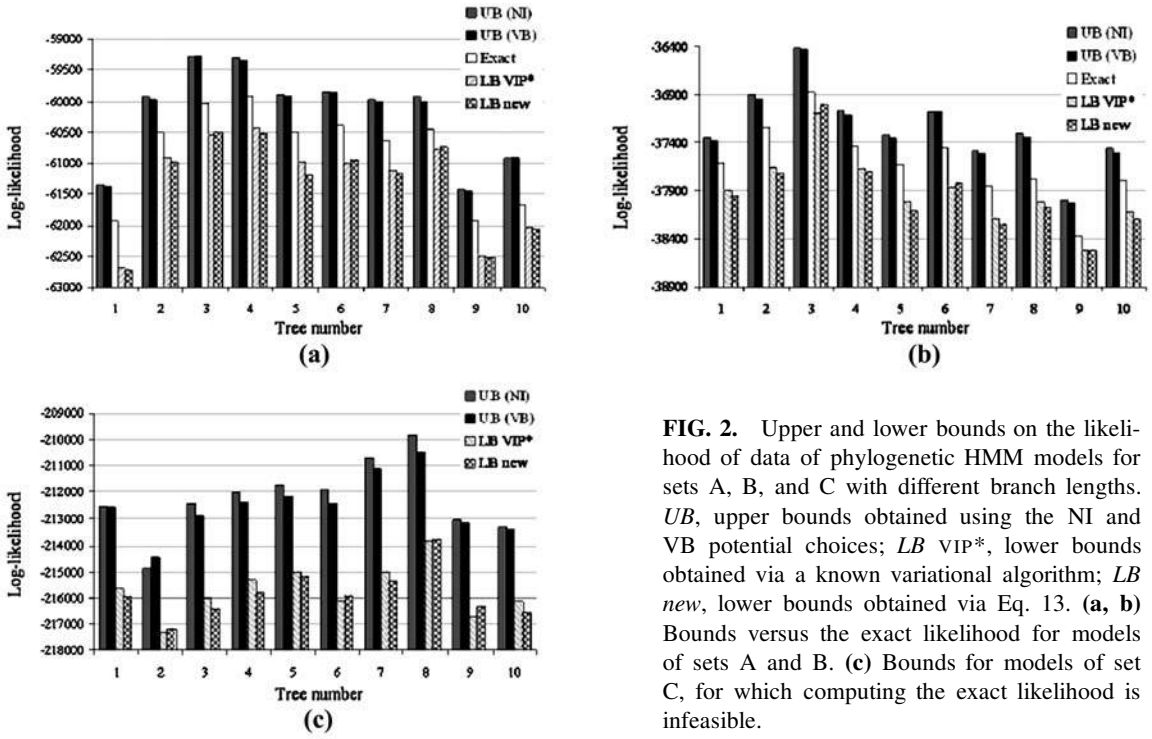

FIG. 2. Upper and lower bounds on the likelihood of data of phylogenetic HMM models for sets A, B, and C with different branch lengths. *UB*, upper bounds obtained using the NI and VB potential choices; *LB* VIP*, lower bounds obtained via a known variational algorithm; *LB new*, lower bounds obtained via Eq. 13. **(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.

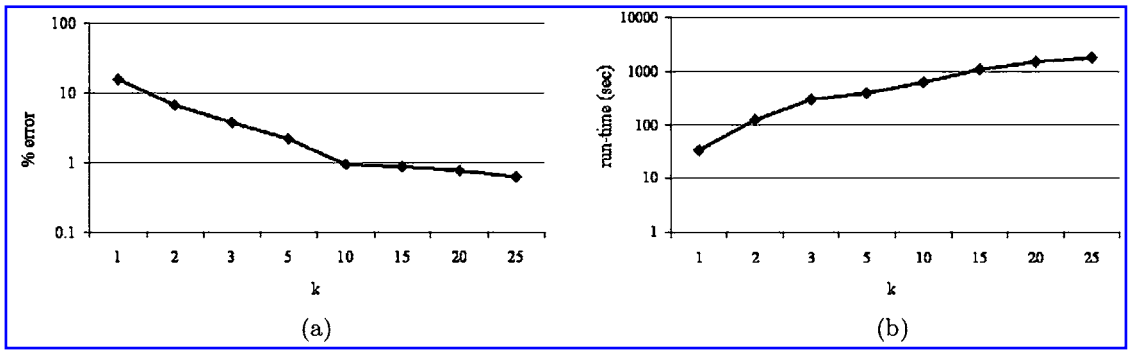

FIG. 3. Accuracy and run-time as a function of parameter k of decomposing the model. **(a)** Accuracy as a function of k. **(b)** Run-time as a function of k.

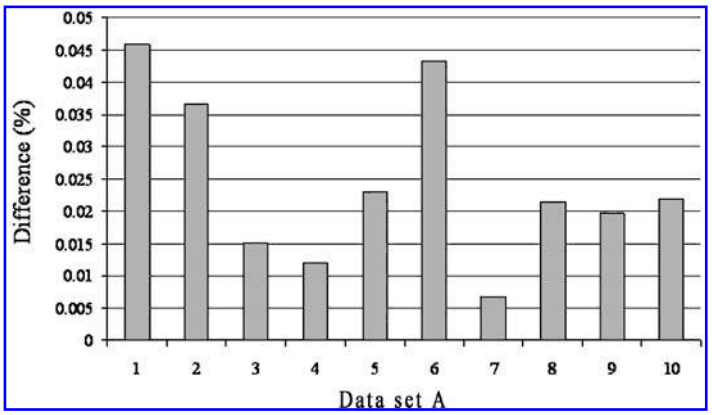

FIG. 4. The difference in accuracy between upper bounds computed via Eq. 5 and via Eq. 6.

compute the exact likelihood, but the small gap between the upper and lower bounds allows us to predict the likelihood with a small error.

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%). In addition, the lower bounds computed via Eq. 13 are similar to the variational bounds computed using VIP* and have almost identical average.

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 that 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 81.25 which equals the average probability table size in the model. 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 bounds are appreciably tight and,

together with other variational methods, allow us 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.

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 classical 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.

The increasing size and complexity of probabilistic models for various applications led an intensive research on approximation methods. Some upper bounds devised for tasks similar to those discussed herein (Wainwright et al., 2005; Globerson and Jaakkola, 2007; Larkin, 2003) may be applied in conjunction with our method, to enhance the tightness by taking a minimum over the results.

6. APPENDIX

Upper bounds on the MPE probability with applications to reconstruction of medical images

Finding the Most Probable Explanation (MPE) assignment to some set of variables H given that another set of variables has been observed $Y = y$ is a query that is common to many applications. For example, in image reconstruction, Y can denote a set of observed pixels of a corrupted image and H may denote a set of pixels in the original image whose values need to be inferred from the observation $Y = y$. In such cases one needs to compute $P(h^*, y) = \max_h P(h, y)$ where h^* is an assignment to H, called the MPE assignment, that maximizes $P(h, y)$. This computation is similar to that of computing the likelihood of evidence $P(y) = \sum_h P(h, y)$ and therefore approximations via upper bounds are desired. In contrary, obtaining lower bounds to $P(h^*, y)$ is immediate because any assignment h' provides the lower bound $P(h', y)$ which can be easily computed.

Several methods for setting upper bounds were previously suggested, among which we mention Detcher & Rish (2003) who iteratively remove dependencies between potentials in the model to create mini-buckets, Weiss, Yanover, & Meltzer (2007) who use linear programming relaxation, and Wainwright, Jaakkola, and Wilsky (2005) where the upper bounds are the result of a convex combination of trees in the model.

Section 3.1 provided upper bounds on the likelihood of evidence y for some of the variables $Y \subseteq X$. We now develop upper bounds also to the probability of the most probable assignment h^* of $H \subseteq X$ when observing $Y = y$. Recall that $P(h^*, y) = \max_h P(y, h) = \max_h \prod_i \Psi_i(d_i)$. Consequently, Eq. 1 can be replaced with

$$P(h^*, y) \leq \max_h \sum_i w_i(h) \left(\Psi_i(d_i)\right)^{(1/w_i(h))} \tag{14}$$

Given a tractable distribution $Q(x) = \prod_{i=1}^n \Phi_i(d_i)$ and setting weights $w_i(h) = \frac{\log \Phi_i(d_i)}{\log Q(h,y)}$, we can upper bound $P(h^*, y)$ similarly to Eq. 3, with the difference of replacing the sum over instantiations h with a maximum over them:

$$P(h^*, y) \leq \max_h \sum_i \frac{\log \Phi_i(d_i)}{\sum_k \log \Phi_k(d_k)} \prod_m \Phi_m(d_m)^{\frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)}} \tag{15}$$

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. 15 is the following:

$$P(h^*, y) \leq \frac{1}{n} \max_h \sum_{i=1}^{n} \log \Phi_i(d_i) \prod_m \frac{\Phi_m(d_m)^{\frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)}}}{[\log \Phi_m(d_m)]^{1/n}} \tag{16}$$

Consequently, the following theorem holds.

Theorem 3 (upper bound on the MPE). *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 $D_i \subseteq X$. Let h^* be an instance of H for which the probability $P(H = h, Y = y)$ is maximum. Then,*

$$P(h^*, Y = y) \leq \frac{1}{n} \sum_i \max_{D_i} \log \Phi_i(d_i) \left[\max_{h \setminus D_i} \prod_m \frac{\Phi_m(d_m)^{\frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)}}}{[\log \Phi_m(d_m)]^{1/n}} \right]. \tag{17}$$

Proof. The proof is immediate from Eq. 16 where we replace the sum over i with the maximum over h, recalling that maximum over a sum is smaller than or equal to sum of maxima, and divide the maximum over h such that first we take maximum over variables in D_i and then over the rest of the variables in H. ■

Also the generalized bounds on likelihood of evidence given by Eq. 7 extend to the MPE probability, by replacing the sum over all the instantiations h of H with maximum over these instantiations:

$$P(h^*, Y = y) \leq \frac{1}{n^l} \sum_{i_1 \ldots i_l} \max_{D_{i_1} \ldots D_{i_l}} \left[\prod_{t=1}^{l} \log \Phi_{i_t}(d_{i_t}) \right] \max_{h \setminus \{D_{i_1} \ldots D_{i_l}\}} \prod_m \frac{\Phi_m(d_m)^{\frac{\log \prod_{t=1}^{l} \Psi_{i_t}(d_{i_t})}{l \cdot \log \prod_{t=1}^{l} \Phi_{i_t}(d_{i_t})}}}{[\log \Phi_m(d_m)]^{l/n}} \tag{18}$$

Almost in all applications finding the probability of the most probable assignment is not sufficient and the actual assignment h^* is needed. To find an assignment with probability close to the maximum we utilize the computations of Eq. 17; Denote the assignments that yield the maxima for the terms $\max_{D_i} \log \Phi_i(d_i) \left[\max_{h \setminus D_i} \prod_m \frac{\Phi_m(d_m)^{\frac{\log \Psi_i(d_i)}{\log \Phi_i(d_i)}}}{[\log \Phi_m(d_m)]^{1/n}} \right]$ by $\tilde{h}_1, \ldots, \tilde{h}_n$. Our heuristic outputs an instantiation $\hat{h} = \max(\tilde{h}_1, \ldots, \tilde{h}_n)$. Other choices that we tried for $\hat{h} = f(\tilde{h}_1, \ldots, \tilde{h}_n)$ such as weighted average have been found less effective.

Below we apply this method to find high probability assignments for several image models, and demonstrate the small gap between the probability of the chosen assignments h^* and the upper bounds provided by Eq. 18, indicating that the assignments found are close to being optimal.

Super-resolution and image reconstruction

Obtaining high-resolution images is important in various applications, as in the example of medical images where seemingly minor details are often helpful to make a correct diagnosis. However, sometime due to technical limitations or cost considerations the resolution of images is not sufficient. Super-resolution is the process of obtaining a high-resolution image with more pixels than the original image, improving the perceived image content compared to that of the low-resolution image (Fig. 5). The super-resolution image reconstruction methods proved to be useful in medical imaging, satellite imaging, and video applications. Synthetic zooming of regions of interest is an important application in surveillance, forensic, scientific, medical, and satellite imaging. For surveillance or forensic purposes, for instance, it is often needed to magnify objects in the scene such as the face of an individual or the license plate of a car.

Freeman, Pasztor, and Carmichael (2000) modeled super-resolution as a pairwise Markov random field, which is an undirected model where the joint distribution over the variables is given as a product of pairwise potentials $P(X = x) = \prod_{i=1}^{n} \psi_i(x_i, x_j)$ and where X_i and X_j are neighboring variables in the model. They used a maximum a-posteriori estimation technique to find an assignment of the variables in the model

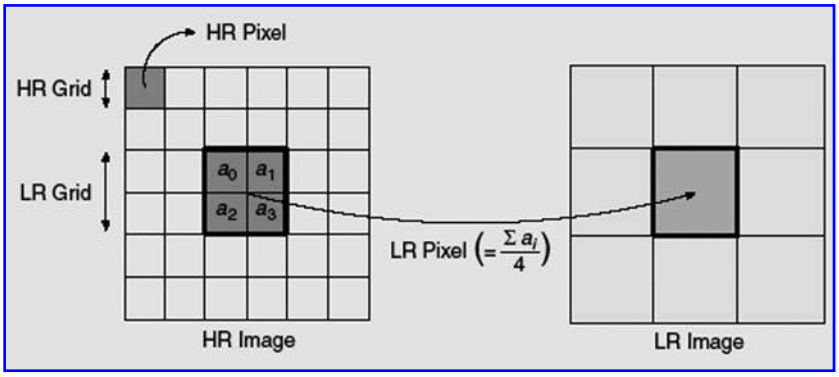

FIG. 5. Super-resolution process where each pixel in the low-resolution image corresponds to four pixels in the high-resolution image.

that yields a sharp image. More specifically, given a low-resolution image y, we are interested in finding the most probable high-resolution scene x, based on probabilities obtained from an adequate training data. For that we divide y into $n \times m$ patches, which may overlap, and each patch in y has its corresponding high-resolution patch in x. When modeled as a graph (Fig. 6), every node Y_{ij} corresponds to a single patch in y, and every node X_{ij} is the respective high-resolution patch. In addition, potentials Ψ are assigned to the edges in the graph and reflect the joint probability for every assignment to a pair of neighboring variables. There are k possible values to each node X_{ij} which are the patches from a training data that best fit the observed patch. Similar to Freeman et al. (2000) we use overlapping patches, and determine the potential between two neighboring patches in x based on their overlap agreement: $\Psi(x_{ij}, x_{lm}) = e^{\beta \cdot d_{lm}^{ij}}$ where d_{lm}^{ij} is the Euclidean distance in the overlap pixels, and β is a user parameter set to 0.001. For simplicity we determine the potential of a patch in y and its corresponding high resolution patch in x similarly by increasing the size of y to that of x and setting: $\Psi(x_{ij}, y_{ij}) = e^{\beta_y \cdot d_y^{ij}}$ where d_y^{ij} is the Euclidean distance between x_{ij} and y_{ij} and β_y is a user parameter set to 0.01.

The goal of finding the most probable assignment to the variables of x is extremely hard when the graph contains more than a few hundred variables. Freeman et al. (2000) used loopy belief propagation (Weiss, 1997) to obtain an assignment and use it to reconstruct their images. Although they show an improvement in the perceived images, it remains a question how close are the assignments they used to the optimal assignments and thus, how much better they could do.

We used the method described here to find upper bounds on the MPE probability of image reconstruction models and used the heuristic method described to find an assignment to the variables X in these models.

We found that our method performs well on images with reduced resolution. In particular we tried the method on images used for medical diagnosis. For example, the low resolution image in Figure 7a

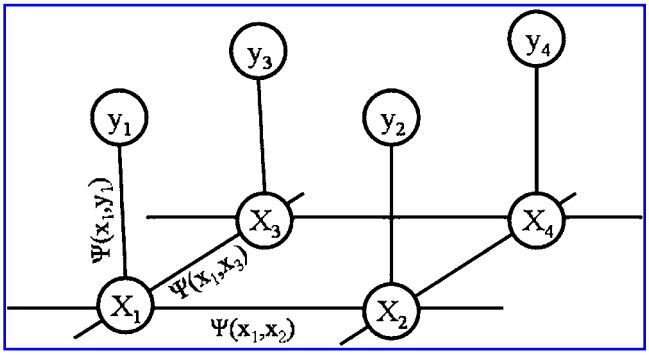

FIG. 6. A Markov network where each node y represents a patch of the observed image, and each node x represents a high-resolution patch of one patch in y. The potentials $|*psi*|(x_i, y_i)$ and $|*psi*|(x_i, x_j)$ indicate the dependencies between nodes.

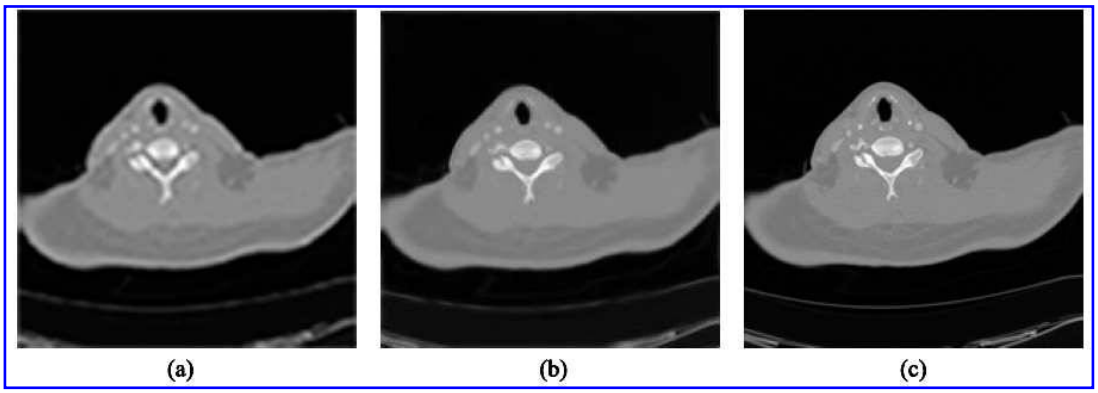

FIG. 7. Computed tomography (CT) image example. **(a)** Low-resolution image. **(b)** High-resolution image obtained by our proposed heuristic. **(c)** Desired high-resolution image.

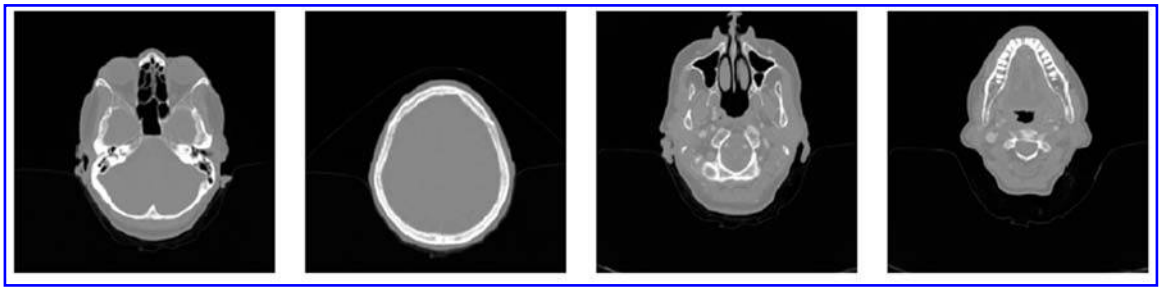

FIG. 8. Example images from a training set of 20 computed tomography (CT) images used to reconstruct the image in Figure 7b.

was modeled using a 200×200 grid. For candidate patches we used 20 CT images from different parts of the body, from which four are shown in Figure 8, and considered three possible values (candidates) for each high resolution node. The corresponding high resolution image retrieved is shown in Figure 7b, and the assignment that generate this image has a log-probability of $-12{,}989$. The upper bound on the log-probability using our method was $-12{,}562$, which is distant 2.6% from the lower bound.

ACKNOWLEDGMENTS

We would like to thank Michael Elad, Ron Kimmel, and Alex Brook for their help and insights related to image reconstruction. The research is supported by the Israel Science Foundation and the Israeli Science Ministry.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Bishop, C., and Winn, J. 2003. Structured variational distributions in VIBES. In: *Artificial Intelligence and Statistics.* Society for Artificial Intelligence and Statistics, Key West, Florida.

Cooper, G. 1990. Probabilistic inference using belief networks is NP-hard. *Artific. Intellig.* 42, 393–405.

Dagum, P., and Luby, M. 1993. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artific. Intellig.* 60, 141–153.

Dechter, R. 1999. Bucket elimination: a unifying framework for reasoning. *Artific. Intellig.* 113, 41–85.

Dechter, R., and Rish, I. 2003. Mini-buckets: a general scheme for bounded inference. *J. ACM* 50, 107–153.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376.

Fishelson, M., and Geiger, D. 2002. Exact genetic linkage computations for general pedigrees. *Bioinformatics* 18, S189–S198.

Freeman, W., Pasztor, E., and Carmichael, O. 2000. Learning low-level vision. *Int. J. Comput. Vision* 40, 25–47.

Geiger, D., Meek, C., and Wexler, Y. 2006. A variational inference procedure allowing internal structure for overlapping clusters and deterministic constraints. *J. Artific. Intellig. Res.* 27, 1–23.

Ghahramani, Z., and Jordan, M. 1997. Factorial hidden Markov models. *Mach. Learn.* 29, 245–273.

Globerson, A., and Jaakkola, T. 2007. Approximate inference using conditional entropy decompositions. *Int. Conf. Artific. Intellig. Statist.*

Jensen, F. 2001. *Bayesian Networks and Decision Graphs.* Springer-Verlag, New York.

Jojic, V., Jojic, N., Meek, C., et al. 2004. Efficient approximations for learning phylogenetic HMM models from data. *Bioinformatics* 20, 161–168.

Jordan, M., Ghahramani, Z., Jaakkola, T., et al. 1999. *An Introduction to Variational Methods for Graphical Models*. MIT Press, Cambridge, MA.

Kschischang, F., Frey, B., and Loeliger, H. 2001. Factor graphs and the sum-product algorithm. *IEEE Trans. Inform. Theor.* 47, 498–519.

Larkin, D. 2003. Approximate decomposition: a method for bounding and estimating probabilistic and deterministic queries. *Uncertainty Artific. Intellig.* 19, 346–353.

Neyman, J. 1971. Molecular studies of evolution: a source of novel statistical problems, 1–27. In: *Statistical Decision Theory and Related Topics*. Academic Press, New York.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Pečarić, J., and Mićić, J. 2005. Some functions reversing the order of positive operators. *J. Linear Algebra Appl.* 396, 175–187.

Saul, L., and Jordan, M. 1996. Exploiting tractable substructures in intractable networks. *Adv. NIPS* 8, 486–492.

Siepel, A., and Haussler, D. 2003. Combining phylogenetic and hidden Markov models in biosequence analysis. *Proc. RECOMB 2003* 277–286.

Wainwright, M., Jaakkola, T., and Willsky, A. 2002. A new class of upper bounds on the log partition function. *Uncertainty Artific. Intellig.* 18, 536–543.

Wainwright, M., Jaakkola, T., and Willsky, A. 2005. Map estimation via agreement on (hyper)trees: message-passing and linear-programming approaches. *IEEE Trans. Inform. Theor.* 51, 3697–3717.

Weiss, Y. 1997. *Belief Propagation and Revision in Networks with Loops. Technical Report 1616.* MIT AI Labs, Cambridge, MA.

Weiss, Y., Yanover, C., and Meltzer, T. 2007. Map estimation, linear programming and belief propagation with convex free energies. *Uncertainty Artific. Intellig.*

Wiegerinck, W. 2000. Variational approximations between mean field theory and the junction tree algorithm. *Uncertainty Artific. Intellig.* 626–633.

Xing, E., Jordan, M., and Russell, S. 2004. Graph partition strategies for generalized mean field inference. *Uncertainty Artific. Intellig.* 602–610.

Address reprint requests to:
Dr. Ydo Wexler
Computer Science Department
Technion
Haifa, 32000, Israel

E-mail: ywex@cs.technion.ac.il