An Approximate Maximum Likelihood Approach, applied to Phylogenetic Trees

Henrik Jönsson[1*] and Bo Söderberg[2]

Complex Systems Division, Department of Theoretical Physics
Lund University, Sölvegatan 14A, S-223 62 Lund, Sweden
http://cbbp.thep.lu.se

(*Journal of Computational Biology* **10**, 737–749 (2003))

Keywords: phylogeny, maximum likelihood, variational, mean-field

Abstract

A novel type of approximation scheme to the maximum likelihood (ML) approach is presented and discussed in the context of phylogenetic tree reconstruction from aligned DNA sequences. It is based on a parameterized approximation to the conditional distribution of hidden variables (related e.g. to the sequences of unobserved branch point ancestors) given the observed data. A modified likelihood, based on the extended data, is then maximized with respect to the parameters of the model as well as to those involved in the approximation. With a suitable form of the approximation the proposed method allows for simpler updating of the parameters, at the cost of an increased parameter count and a slight decrease in performance. The method is tested on phylogenetic tree reconstruction from artificially generated sequences, and its performance is compared to that of ML, showing that the approach is competitive for reasonably similar sequences. The method is also applied to real DNA-sequences from primates, yielding a result consistent with those obtained by other standard algorithms.

[1] Henrik.Jonsson@thep.lu.se, Phone:+46 46 222 3494, Fax:+46 46 222 9686
[2] Bo.Soderberg@thep.lu.se

1 Introduction

Several different types of algorithms have been proposed for inferring phylogenetic trees from sequence data of species (see e.g. Swafford and Olsen 1996; Nei 1996). The theoretically most appealing of these are based on the maximum likelihood (ML) approach, introduced in this context by Felsenstein (1981), and implemented into standard computer packages (Felsenstein 1993; Yang 1997). In ML, a more or less simple, stochastic model is assumed for sequence branching and independent site evolution, resulting in a tree graph expressing the phylogenetic relationship between the observed species. The topology (structure) and geometry (arc lengths) of the tree, and possible additional model parameters, are to be chosen so as to maximize the *likelihood*, defined to be proportional to the probability of the model to produce the observed sequences.

In practice, the maximization of likelihood with respect to the model parameters is a complicated task. Typically, it is done by optimizing one parameter at a time while keeping the others fixed. This is repeated until some criterion for convergence is met. However, even with the simplest evolution models, each single-parameter optimization step requires an iterative procedure, and the method can be quite time-consuming.

In this article, an alternative approach is proposed, where the observed data is extended by means of optimizing a simple parameterized approximation to the conditional probability distributions of the unobserved branch-node sequences. The conventional likelihood L is replaced by an alternative likelihood $\hat{L} < L$, associated with the extended data. The maximization of L with respect to the model parameters is then replaced by the maximization of $\hat{L}$ with respect to these as well as to the parameters of the approximation.

Depending on the complexity of the form of the approximation, this method will decrease performance to some degree, as measured by the achieved value of L, and introduce additional parameters to optimize. This is compensated for by allowing for a simpler updating of model parameters. With a suitable form for the parameterized approximation, also its parameters allow for a simple updating scheme.

The method contains elements both from ML and from the variational approach (Feynman 1972), as used in statistical physics, and will be referred to as a *variational maximum likelihood* approach (*VML*).

2 Background

Before presenting the VML approach in some detail in the nect section, we will here introduce notation, and give some theoretical background. We will briefly discuss independent-site Markov models for the evolution and the associated maximum likelihood approach. We will also briefly review variational methods as used in statistical physics, in particular the specific example given by the mean-field (**MF**) approximation (Parisi 1988).

2.1 Stochastic Independent-Site Mutation Models

We will consider a class of simple stochastic models for the evolution of DNA sequences, where individual sites in the sequence mutate at random according to an identical model, but independently from each other; deletions and insertions are neglected. Speciation is assumed to occur in the form of branching events, where a species bifurcates in two,[3] which continue to evolve independently.

In such a model, a specific set of observed species is assumed to have evolved from a common ancestor, by means of successive bifurcations, and subsequent periods of independent random mutations. Their phylogenetic relationship takes the form of a phylogenetic tree, with the observed species at the leaves, while each branch point corresponds to an unobserved ancestor subject to a speciation event.

There is a priori no reason to assume that the mutation rate is the same in different branches, nor constant in time on a single branch. If it were, one would expect constraints between the evolutionary distances along the different branches in the tree. This also makes the position in the tree of the common ancestor ambiguous, in particular if the assumed model is invariant under time reversal.

Thus, one is led to consider unrooted trees without a definite temporal ordering along branches. Assuming interior nodes of order three, a tree with N leaves will have $N-2$ internal nodes and $2N-3$ links, and allow for $(2N-5)!! = (2N-5)\times(2N-7)\times\ldots\times 3\times 1$ distinct leaf-labeled topologies. The set of nodes (vertices) will be denoted by $\mathbb{V} = \mathbb{V}_1 \cup \mathbb{V}_3$, decomposing into the set of leaves, $\mathbb{V}_1$, and the set of branch points, $\mathbb{V}_3$; the set of links (edges) will be denoted by $\mathbb{E}$. Figure 1 shows a schematic phylogenetic tree for five observed species, connected via three unobservable ancestors.

Consider a set of N aligned homologous DNA sequences, corresponding to a set of N species. Then at each site in the sequences, a combination $S = (s_1, \ldots, s_N)$ is observed, where each s_i is a symbol in the alphabet $\mathcal{A} = \{A, C, G, T\}$ of size $K = 4$. For a

[3] Multifurcations can be seen as sequences of several bifurcations

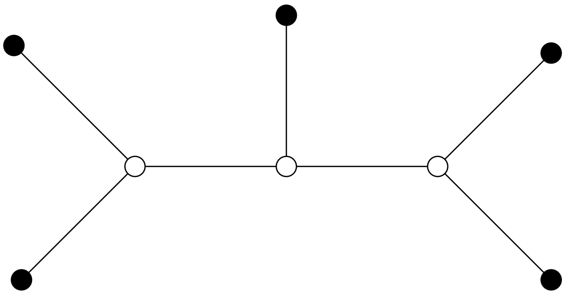

Figure 1: A phylogenetic tree.

given tree topology, the set of $N-2$ branch point species corresponds to an unknown combination $I = (s_{N+1}, \ldots, s_{2N-2}) \in \mathcal{A}^{N-2}$ at the same site. In what follows, sums or products over S will be understood as running over $S \in \mathcal{A}^N$; likewise, those over I will be understood to run over $I \in \mathcal{A}^{N-2}$.

In a model of the above described type, the probability for a definite aggregate symbol combination $SI = (s_1, \ldots, s_{2N-2}) \in \mathcal{A}^{2N-2}$ is given by a product of single link factors,

$$P_{SI} = \prod_{\{k,l\} \in \mathbb{E}} T^{\{k,l\}}_{s_k, s_l}, \tag{1}$$

where $\{k, l\}$ denotes an edge connecting nodes k and l. The entries in a specific link factor $T^{\{k,l\}}$ depend on the specific mutation model, as well as on the link-specific parameters, such as evolutionary distance. Different models put more or less severe constraints on the link factors.

With a given topology, and with given link factors defining P_{SI}, the probability of the observed single-site combination S becomes

$$P_S = \sum_I P_{SI}. \tag{2}$$

2.2 Maximum Likelihood

In a situation where aligned homologous sequences from a set of N species are given, each individual site can be viewed as an independent experiment, the outcome of which is the ordered set S of observed symbols of the considered species at that particular site. The observation of the full sequences, assumed to have a common length M, thus corresponds to repeating the same experiment M times and gathering statistics of the outcome. The statistical information can be collected into an *observed distribution* Q

over the single-site observations S,

$$Q_S = \frac{M_S}{M}, \tag{3}$$

where M_S is the number of sites where the single-site combination S appears. Obviously, we have $\sum_S Q_S = 1$.

To these experimental data, the parameters of a model is to be fitted. Assuming the model to yield the probability P_S for a single-site combination S, the probability that the model produce the observed multiplicities M_S is given by the product $\prod_S P_S^{M_S}$ over all single-site combinations $S \in \mathcal{A}^N$, multiplied by the proper combinatorial factor $M!/\prod_S M_S!$, that takes into account the number of distinct ways to realize the observed multiplicities (by permuting sites). In the limit of very long sequences the combinatorial factor is dominated by $\prod_S Q_S^{-M_S}$, which we use to define a suitably normalized likelihood as

$$L = \prod_S \left(\frac{P_S}{Q_S}\right)^{MQ_S}, \tag{4}$$

associated with a model yielding the probabilities P, given the observed distribution Q.

Note that the approximation of the combinatorial factor only affects the normalization of the likelihood in a way that is independent of P, and so is harmless. Although somewhat unconventional, this normalization is very natural since it gives a unit likelihood for a perfect fit, i.e. for $P = Q$.

Maximizing L corresponds to minimizing its *negative logarithm* divided by the sequence length M, to be referred to as the *free energy* per site, F,

$$F = \sum_S Q_S \log\left(\frac{Q_S}{P_S}\right) \geq 0. \tag{5}$$

In terms of F, the likelihood is given by $L = \exp(-MF)$. The free energy is related to the *mutual entropy* between Q and P; it is a *strictly convex, non-negative* function of the probabilities P_S, and hence of P_{SI}, vanishing only if P and Q are identical. For the case of a parameterized model for P, the convexity of F is no guarantee for a unique global likelihood maximum (Chor et al. 2000). Thus, if a model is capable of producing a P that exactly matches Q it will yield a vanishing F; otherwise a strictly positive F will result.

2.3 Variational Method - General Principles

Consider a situation where a complicated theoretical probability distribution P over a set of variables S is given, and for simplicity one wishes to approximate it in an optimal way by a simpler parameterized expression V of a certain form.

In the *variational approach* (Feynman 1972), one considers an associated *variational free energy* G, defined by

$$G = \sum_S V_S \log\left(\frac{V_S}{P_S}\right), \tag{6}$$

which is to be minimized with respect to the parameters in V. Note that G is a non-negative convex function of V, with a unique vanishing minimum for $V = P$.

G has a very similar appearance to the free energy F of the ML approach, eq. (5). Note however the changed role of P: In eq. (5), P represents a parameterized model to be fitted to Q, while in eq. (6), P defines a given target distribution to which the parameterized approximation V is to be fitted.

2.4 The Mean-Field Approximation

The distribution over symbols considered in the reconstruction of phylogenetic trees is highly analogous to the thermal distributions encountered in the statistical physics of spin systems.

In the context of spin systems, a common application of the variational approach is the *Mean-Field* (**MF**) approximation (Parisi 1988), where a given spin distribution is approximated by one that is factorized over the distinct spin variables. Thus, for a set of K-state spins, $S = (s_1 \ldots s_N) \in \mathcal{A}^N$, a given distribution P_S is to be approximated in an optimal way by a factorized distribution $V_S = \prod_{k=1}^N v_{s_k}^{(k)}$, as defined by the minimization of the free energy

$$G(v) = \sum_{k=1}^{N} \sum_{s \in \mathcal{A}} v_s^{(k)} \log v_s^{(k)} - \sum_S \log P_S \prod_{k=1}^{N} v_{s_k}^{(k)}. \tag{7}$$

The expression $-\log P_S$ can be interpreted as a cost function (or Hamiltonian) H_S, in terms of which G can be written as

$$G(v) = \sum_{k=1}^{N} \sum_{s \in \mathcal{A}} v_s^{(k)} \log v_s^{(k)} + \langle H_S \rangle_v \,, \tag{8}$$

where the first term is the negative of the entropy, while the second expresses the average cost. Minimization of G with respect to each single-spin distribution $v^{(k)}$ yields the *MF equations*,

$$v_s^{(k)} \propto \exp\left(-\partial \langle H \rangle_v / \partial v_s^{(k)}\right), \tag{9}$$

where the constant of proportionality is fixed by the normalization, $\sum_{s\in\mathcal{A}} v_s^{(k)} = 1$. (Locally) optimal distributions can be found by an iterative updating scheme based on eq. (9).

3 Variational Maximum Likelihood

We are now prepared to formulate the hybrid approach VML, where an approximation to the likelihood is maximized, in the context of phylogenetic tree reconstruction. It contains elements both of ML and of the variational approach.

3.1 General Idea

When applying ML to a stochastic mutation model as described in section 2.1, the factorization in P_{SI} cannot be fully exploited, since it is lost in P_S due to the summation over the hidden symbols I.

If, in addition to the observed sequences at the leaves of the tree, also the corresponding sequences of the unknown ancestors at branchpoints were known (Friedman et al. 2002), one could apply ML to the extended distribution Q_{SI}, corresponding to minimizing a modified free energy $\hat{F}$,

$$\hat{F} = \sum_{SI} Q_{SI} \log\left(\frac{Q_{SI}}{P_{SI}}\right) . \tag{10}$$

This would simplify the maximization of likelihood considerably due to the factorization of P, corresponding to a decomposition of $\hat{F}$ as the sum of terms, each associated with a single link. As a result, the parameters for a single link would be determined by the minimization of an expression like

$$\sum_{i,j\in\mathcal{A}} q_{ij} \log \frac{q_{ij}}{p_{ij}}, \tag{11}$$

where i, j correspond to the respective symbols at the two nodes connected by the link, while q and p denote their joint marginal distribution as derived from Q_{SI} and P_{SI} respectively.

To take advantage of the factorization property of P_{SI}, we now propose an approximative approach, generically described as follows.

The generic VML method:

- Choose a parameterized expression for the conditional distribution $V_{I|S}$ of hidden symbols, given the observed ones.
- Q_{SI} is now determined as $Q_S V_{I|S}$. Consider the corresponding free energy $\hat{F}$, as given by eq. (10).
- The minimization of $\hat{F}$ with respect to the model parameters in P is straightforward, yielding an optimal value of $\hat{F}$, associated with this particular $V_{I|S}$.
- The resulting free energy should then be minimized also with respect to the parameters of $V_{I|S}$.

Thus, in VML, $\hat{F}$ is to be minimized both with respect to the model parameters defining P_{SI}, and the extra parameters defining $V_{I|S}$.

Note that $\hat{F}$ approximates F from above, which can be seen by rewriting it as

$$\hat{F} = F + \sum_S Q_S \left[\sum_I V_{I|S} \log \left(\frac{V_{I|S}}{P_{I|S}} \right) \right] \geq F. \tag{12}$$

It has the obvious form

$$\hat{F} = F + \sum_S Q_S G_S, \tag{13}$$

where for each S, G_S can be interpreted as a *variational free energy* for the approximation of $P_{I|S}$ by $V_{I|S}$.

With a sufficiently general form for $V_{I|S}$, it would match $P_{I|S}$ at optimality, making the second term above vanish. Thus, in such a case, $\min \hat{F} = \min F$. The resulting approach would be an exact reformulation of conventional ML.

3.2 Factorized VML

Here we will consider a particularly simple implementation of VML, by employing a particular form for the extension $V_{I|S}$, constrained to be *factorized* over the internal nodes,

$$V_{I|S} = \prod_{k \in \mathbb{V}_3} v^{(k)}_{i_k|S}. \tag{14}$$

Then, the modified free energy $\hat{F}$ can be simplified to read

$$\hat{F} = \sum_S Q_S \log Q_S + \sum_S Q_S \sum_{k \in \mathbb{V}_3} \sum_{i \in \mathcal{A}} v^{(k)}_{i|S} \log v^{(k)}_{i|S} - \sum_S Q_S \sum_{\{k,l\} \in \mathbb{E}} \sum_{i,j \in \mathcal{A}} v^{(k)}_{i|S} v^{(l)}_{j|S} \log T^{\{k,l\}}_{ij}, \tag{15}$$

where $\{k,l\}$ labels a link connecting two nodes k,l, while $T^{\{k,l\}}$ is the corresponding link factor. If k (or l) refers to an external node, $v^{(k)}_{i|S}$ is to be interpreted as δ_{i,s_k}.

For each S, the associated internal node distributions $\{v^{(k)}_{i|S}\}$ minimize

$$\sum_{k\in\mathbb{V}_3}\sum_{i\in\mathcal{A}} v_i^{(k)}\log v_i^{(k)} - \sum_{\{k,l\}\in\mathbb{E}}\sum_{i,j\in\mathcal{A}} v_i^{(k)}v_j^{(l)}\log T_{ij}^{\{k,l\}}, \tag{16}$$

where the "$|S$" has been stripped off for clarity. This has the precise form of a variational free energy for the MF approximation, cf. eq. (8). The condition for a local minimum of $\hat{F}$ with respect to $v^{(k)}$ yields the MF equations (9), which in this case can be written as

$$v^{(k)}_{i|S} \propto \prod_{l\in\mathcal{N}_k}\prod_{j\in\mathcal{A}}\left(T_{ij}^{\{k,l\}}\right)^{v^{(l)}_{j|S}}, \tag{17}$$

normalized such that $\sum_{i\in\mathcal{A}} v_i^{(k)} = 1$. Here, $\mathcal{N}_k$ stands for the set of nodes that are neighbors to k. Eq. (17) can be used for iteratively updating v.

There is an obvious ambiguity in the link factors, associated with factors depending on a single internal symbol – such factors can be exchanged between the link factors associated with the three links surrounding it, without affecting their product. For a fixed link $\{k,l\}$, we can use this freedom to force the corresponding link factor $T^{\{k,l\}}$ to equal the marginal two-symbol distribution $p^{\{k,l\}}$, derived from P_{IS} by summing over the remaining nodes.

Then the part of $\hat{F}$ relevant for a link factor $p^{\{k,l\}}$ can be written in the form of eq. (11). This means that $p^{\{k,l\}}$ should be chosen to optimally fit (in the ML sense) the corresponding marginal two-symbol distribution $q^{\{k,l\}}$, as determined by Q_{IS}. For a parameter a in $p^{\{k,l\}}$, optimality thus implies

$$\frac{\partial\hat{F}}{\partial a} \equiv \sum_{i,j\in\mathcal{A}} \frac{q_{ij}^{\{k,l\}}\partial p_{ij}^{\{k,l\}}/\partial a}{p_{ij}^{\{k,l\}}} = 0, \tag{18}$$

if the optimal value of a is in the interior of its allowed interval.

One might fear that the optimization of the link factors in the form of pair-distributions might yield inconsistent results for two neighboring links, since both pair-distributions determine the marginal distribution $p^{(k)}$ for their common node k. This is no problem – both are consistent with an identical $p^{(k)}$, minimizing (if not fixed by the model) its own relevant part of $\hat{F}$, given by

$$\sum_{i\in\mathcal{A}} q_i^{(k)}\log\left(\frac{q_i^{(k)}}{p_i^{(k)}}\right). \tag{19}$$

4 Application to the JC model

While the above considerations are somewhat abstract, we will in this section be very concrete and give a detailed description for one of the simplest mutation models, the Jukes-Cantor **JC** model (Jukes and Cantor 1969). It is highly constrained: For an arbitrary alphabet size K, a link factor (as above taken as the corresponding marginal pair probability) must take the form

$$p_{ij} = \frac{1}{K^2}\left(1 - a + Ka\delta_{ij}\right), \tag{20}$$

with a single free parameter $a \in [0, 1]$ per link, given by $a = \exp(-t)$, with t an associated evolutionary distance. For each single node k, this yields a uniform marginal single-symbol distribution, $p_i^{(k)} = 1/K$. For a given topology, the full P_{SI} becomes

$$P_{SI} = \frac{1}{K} \prod_{\{k,l\}\in\mathbb{E}} \left(\frac{(1 - a_{\{k,l\}})}{K} + a_{\{k,l\}}\delta_{ij} \right). \tag{21}$$

4.1 ML approach for JC

In the ML approach one would for the JC model consider the free energy F of eq. (5), with P_S as given by summing P_{SI} in eq. (21) over I. F is to be minimized both with respect to the topology, as given by the structure of the tree, and with respect to its geometry, as defined by the link parameters a.

Thus, for a fixed topology, the link parameters $\{a\}$ are to be chosen so as to minimize

$$F = \sum_S Q_S \log(Q_S) - \sum_S Q_S \log\left(\sum_I P_{SI}(\{a\})\right). \tag{22}$$

For the update of a single link parameter $a = a_{\{k,l\}}$, it is advantageous to write F as

$$\begin{aligned} F &= \text{const.} - \sum_S Q_S \log\left(\sum_{i,j\in\mathcal{A}} u_i^{(S)} \left(\frac{1-a}{K} + a\delta_{ij}\right) w_j^{(S)}\right) && (23) \\ &= \text{const.} - \sum_S Q_S \log\left(\frac{1-a}{K} \sum_{i\in\mathcal{A}} u_i^{(S)} \sum_{j\in\mathcal{A}} w_j^{(S)} + a \sum_{i\in\mathcal{A}} u_i^{(S)} w_i^{(S)}\right), && (24) \end{aligned}$$

where $u_i^{(S)}$ and $w_j^{(S)}$ represent probabilities associated with the observed symbols S in the two subtrees joined by the link $\{k, l\}$, conditional upon fixed symbols i, j at the nodes

k, l attached to the link. They do not depend on a, and differentiation of F with respect to a yields

$$\frac{\partial F}{\partial a} = -\sum_S \frac{Q_S}{a - z_S}, \tag{25}$$

which should vanish at minimum. This expression has singularities (simple poles) at S-dependent positions z_S, given by

$$z_S = -\frac{\sum_{i\in\mathcal{A}} u_i^{(S)} \sum_{j\in\mathcal{A}} w_j^{(S)}}{K \sum_{i\in\mathcal{A}} u_i^{(S)} w_i^{(S)} - \sum_{i\in\mathcal{A}} u_i^{(S)} \sum_{j\in\mathcal{A}} w_j^{(S)}}, \tag{26}$$

guaranteed to lie outside the interval $[-1/(K-1), 1]$.

Thus, for a in the physical interval $[0, 1]$, $\partial F/\partial a$ is strictly increasing; if it has a zero in this interval, this defines the optimal value of a (for fixed values of the other parameters); otherwise it is positive on the whole interval $[0, 1]$, in which case $a = 0$ is the optimal value, or negative, in which case $a = 1$ is optimal. An optimum in the interior of $[0, 1]$ has to be found by some iterative method, such as Newton-Raphson, or binary search.

In this way, one parameter at a time can be locally optimized for fixed values of the others, eventually leading to a (local) minimum of F for the chosen topology; this value is taken as a measure of F for that topology.

The optimization with respect to tree topology can be done in different ways. To strictly ensure that the best topology is found, a search of all possible topologies must be performed. Often, though, one settles for a neighborhood search, where an initial topology is chosen at random, or better, by means of some heuristic, and an optimization if performed with respect to link parameters, yielding a value of F for the chosen topology. Then, neighboring topologies, obtained e.g. by rearranging the tree around one of the shorter links (i.e. one with a large a), are investigated. If one of these yields a lower F, it is chosen as the new present topology, and its neighbours are checked. When no more improvements can be made in this way, the present topology is considered optimal. For a more detailed discussion of heuristic topology optimization, see e.g. (Swafford and Olsen 1996).

4.2 Factorized VML approach for JC

For the JC model, the VML method with a factorized V becomes particularly simple. The extended free energy $\hat{F}$ of eq.(15) simplifies to

$$\hat{F} \;=\; \sum_S Q_S \log Q_S + \sum_S Q_S \sum_{k\in\mathbb{V}_3} \sum_{i\in\mathcal{A}} v_{i|S}^{(k)} \log v_{i|S}^{(k)} + \log(K) - \tag{27}$$

$$- \sum_{\{k,l\}\in\mathbb{E}} \left\{ \log\left(\frac{1-a_{\{k,l\}}}{K}\right) + \log\left(\frac{1+(K-1)a_{\{k,l\}}}{1-a_{\{k,l\}}}\right) \sum_S Q_S \sum_{i\in\mathcal{A}} v^{(k)}_{i|S} v^{(l)}_{i|S} \right\} .$$

The part relevant for the conditional distribution $v^{(k)}_{i|S}$ of the symbol i at a specific branch point k for a fixed observed symbol combination S then reads

$$\sum_{i\in\mathcal{A}} v^{(k)}_{i|S} \log v^{(k)}_{i|S} - \sum_{l\in\mathcal{N}_k} \log\left(\frac{1+(K-1)a_{\{k,l\}}}{1-a_{\{k,l\}}}\right) \sum_{i\in\mathcal{A}} v^{(k)}_{i|S} v^{(l)}_{i|S}, \tag{28}$$

where, as before, $v^{(l)}_{j|S}$ is to be interpreted as δ_{j,s_l} whenever l happens to be an external node. The resulting optimality condition, eq. (17), hence becomes

$$v^{(k)}_{i|S} \propto \prod_{l\in\mathcal{N}_k} \left(\frac{1+(K-1)a_{\{k,l\}}}{1-a_{\{k,l\}}}\right)^{v^{(l)}_{i|S}} , \tag{29}$$

with the constant of proportionality fixed by the normalization condition $\sum_{i\in\mathcal{A}} v^{(k)}_{i|S} = 1$. Eq. (29) can be used to iteratively update the parameters of the approximation $V_{I|S}$.

As to the link parameters, their optimization becomes considerably simplified in VML as compared to the case with ML. The part of $\hat{F}$ that is relevant for a specific link parameter $a_{\{k,l\}}$ comes entirely from the last term in eq. (27), and reads

$$-\log\left(\frac{1-a_{\{k,l\}}}{K}\right) - \log\left(\frac{1+(K-1)a_{\{k,l\}}}{1-a_{\{k,l\}}}\right) \sum_S Q_S \sum_{i\in\mathcal{A}} v^{(k)}_{i|S} v^{(l)}_{i|S}, \tag{30}$$

where the double sum in the last term can be interpreted as $O_{\{k,l\}} = \left\langle \mathbf{v}^{(k)} \cdot \mathbf{v}^{(l)} \right\rangle_Q$ in obvious notation, i.e. as a weighted average of the overlap between the distributions associated with the link's endpoints. Demanding a vanishing derivative with respect to $a_{\{k,l\}}$ directly yields the optimal value in explicit form as

$$a_{\{k,l\}} = \frac{K \left\langle \mathbf{v}^{(k)} \cdot \mathbf{v}^{(l)} \right\rangle_Q - 1}{K-1}, \tag{31}$$

which is automatically in the interval $[-\frac{1}{(K-1)}, 1]$, since the overlap must satisfy $O_{\{k,l\}} \in [0,1]$. This is very reasonable: A maximal overlap $O = 1$ yields $a = 1$, while an overlap of $O = 1/K$, corresponding to complete lack of correlation, yields $a = 0$; the unphysical case of a vanishing overlap, $O = 0$, corresponding to maximal anticorrelation, yields $a = -1/(K-1)$. Eq. (31), with negative values adjusted to zero, can be used for updating the parameters $\{a_{\{k,l\}}\}$ of the model distribution P_{SI}.

In this way, one parameter at a time can be locally optimized for fixed values of the others, guaranteeing the decrease of $\hat{F}$, and eventually leading to a (local) minimum of $\hat{F}$ for the chosen topology; this value is taken as a measure of $\hat{F}$ for that topology. The optimization with respect to topology can be performed in the same way as in ML.

4.3 Note on more general models

The JC model is the simplest random independent site mutation model. It has been extended by allowing for differentiation in transition rates, as well as for different single-node probabilities (Kimura 1980; Tamura and Nei 1993; Felsenstein 1981; Hasegawa et al. 1985). Also more general models generated from arbitrary rate matrices (reversible or not) have been studied (Yang 1994; Gu and Li 1996).

In analogy to the proper ML approach, VML (with a factorized or more general $V_{I|S}$) can be applied equally well to any of these models. The actual updating equations will of course change, but will still be based on eqs. (17, 18).

5 Numerical Explorations

In this section we present the results of some simple computer experiments to gauge the performance of VML by comparing it to standard ML. We have consistently used JC for transition probabilities, both when generating sequences for the testbed problems and as the underlying model in the algorithms used to infer the trees.

First, we have probed the method for a homogenous (all a equal) tree with infinite sequence length, and varying a. In the next test we used randomly generated trees, where we compared the achieved values for F obtained by the methods when given the correct (generated) tree topology, and also checked whether the tree with correct topology has the lowest F. Finally we used DNA-sequences from primates to check whether the preferred topology from VML is the one achieved by standard methods.

5.1 Infinite Sequence Test

If random sequences are generated, e.g. according to the JC model, for a tree with given topology and geometry, the single-site distributions Q_S will approach the corresponding model probabilities P_S in the limit of infinitely long sequences. Thus, for each of the K^N possible S, the JC model yields

$$Q_S = \sum_I \frac{1}{K} \prod_{\{k,l\}\in\mathbb{E}} \left(\frac{1-a_{\{k,l\}}}{K} + a_{\{k,l\}} \delta_{i_k,i_l} \right) . \tag{32}$$

Given these data and the correct topology, we expect the ML algorithm to be able to produce the correct link lengths and a vanishing F within numerical limitations. Being based on maximizing an approximation to the real likelihood, the VML algorithm can be expected to perform slightly worse on such data.

We have probed VML and ML with infinite-sequence data for four species[4], based on a tree where the five link parameters were all set equal to a common value a_0. Figure 2 A shows the resulting values of F as a function of the input parameter a_0. When applied to the correct topology, ML clearly gives an essentially perfect fit as expected. VML is seen to perform well for short links ($a_0 \approx 1$), and slightly worse for longer links (smaller a_0), though still not far from optimal. For an incorrect topology both algorithms produced F values well above the ones achieved for the correct topology, showing that either algorithm identifies the correct topology. The resulting individual link parameters were essentially exactly a_0 for the ML algorithm when probed on the correct topology. For VML the resulting link parameters tend to deviate somewhat for cases with smaller a_0 as shown in figure 2 B.

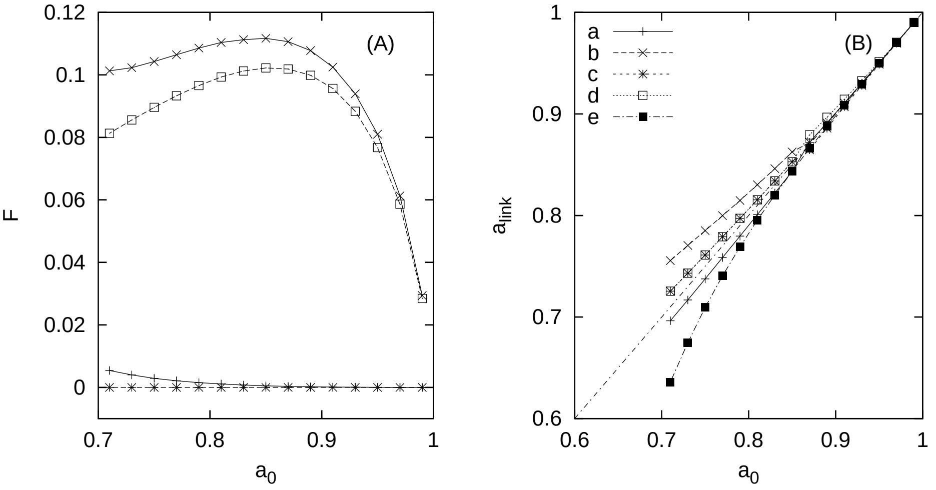

Figure 2: Infinite sequence results for N=4. A) The free energy per site (F) versus link lengths of generated tree (a_0) plotted for ML (correct topology ($*$), wrong topology (□)) and VML (correct topology ($+$), wrong topology ($\times$)). B) Link lengths resulting from the VML algorithm applied on the correct topology (a_{link}) versus link lengths of generated tree (a_0). Links a and b connects one pair of species, c and d the other pair, and e is the link between the internal nodes.

5.2 Tests with Artificial Sequences

We have used testbeds with artificial sequences generated according to the JC model on random trees for varying numbers of species, N, and sequence lengths, M. The random trees were defined by first generating a random topology, then setting link parameters

[4] For three species the mean-field approximation becomes exact.

according to a specific probability distribution. Finally, sequences were generated based on JC on this tree.

A random tree topology is defined by starting with two nodes connected by a single link. Then a new node is connected to an existing link chosen at random; this is repeated until there are N external nodes. This results in an equal probability for each possible leaf-labeled topology. Link parameters are independently generated as $a := R^{t_0}$, with R a distinct uniform random number in $[0, 1]$, while t_0 is a common parameter, setting the temporal scale of the generated links. The motivation for using this probability distribution is that the time associated with a link will follow an exponential distribution, corresponding to a constant branching rate. For t_0 we have used values between 0.05 and 0.3, corresponding to an average a between 0.95 and 0.77. The a-distributions are shown in figure 3. Finally, sequences are generated randomly according to the JC model.

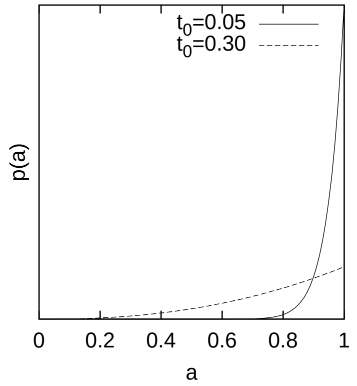

Figure 3: Link parameter distributions ($p(a)$) for the studied test sets.

5.2.1 Inferring the Geometry

The ML and VML algorithms were used to infer the geometry of each tree given the correct topology and the external sequences, and the free energy, F, was used as a measure of the quality of the obtained geometry. We have probed different numbers of species, $N = \{4, 8, 12, 16\}$, link distributions, $t_0 = \{0.05, 0.1, 0.3\}$, and sequence lengths, $M = \{250, 500, 1000\}$, and constantly used the alphabet size $K = 4$ for the sequences.

In figure 4 the achieved F values are plotted versus N for different values of M and t_0. As can be seen in the figure, the difference between ML and VML is hardly noticable. However there is a small difference, as shown in figure 5.

As in the case with infinite sequence lengths we can again see that the difference increases for trees where longer links are used. A larger M results in lower values of F for both algorithms, and the difference has a quite small M-dependence.

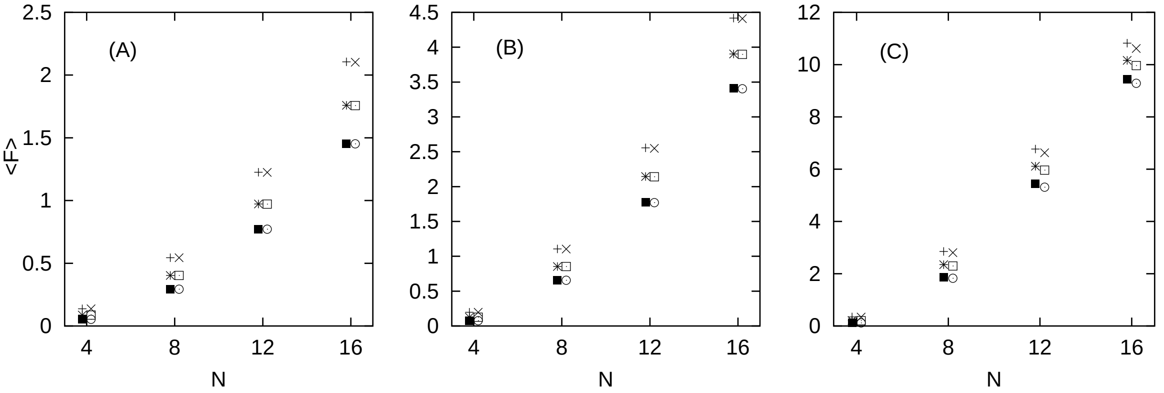

Figure 4: Average of free energy per site $< F >$ versus number of species N for ML and VML. Each data point is the average of 100 randomly generated trees. $K = 4$ for all runs. The plots shows $M = 250$ (ML($\times$) and VML($+$)), $M = 500$ (ML($\square$),VML($*$)) and $M = 1000$ (ML($\circ$),VML($\blacksquare$)). Link parameters are generated using $a := R^{t_0}$ where R is a uniform random number in [0,1]. A) $t_0 = 0.05$, B) $t_0 = 0.1$ and C) $t_0 = 0.3$.

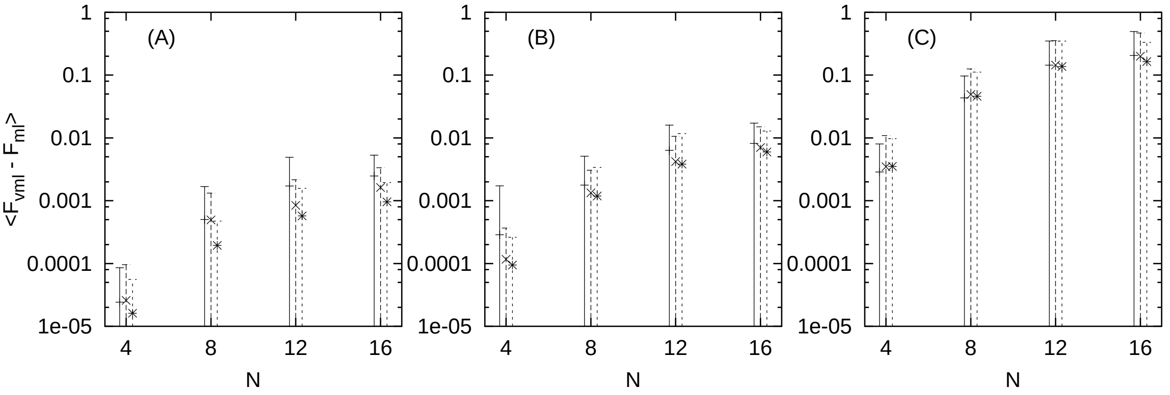

Figure 5: The average of the difference in F versus N for the same data set as in figure 4. The plots show the difference for $M = 250$ ($+$), $M = 500$ ($\times$) and $M = 1000$ ($*$). The errorbars indicate the standard deviation. A) $t_0 = 0.05$, B) $t_0 = 0.1$ and C) $t_0 = 0.3$.

5.2.2 Topology Inference

We also tested the ability to identify the correct topology, in the sense that it yields the lowest free energy compared to other topologies. The tests were performed on trees with four species, and link parameters were generated as above, using t_0 between 0.05 and 0.3. Sequences of length $M = \{250, 500, 1000\}$ and $K = 4$ were used.

The fraction, f_u, of trees where the correct topology does not yield the lowest free energy was measured for ML and VML, and the result is shown in table 1. For both algorithms, longer sequences result in a lower f_u, as expected. VML seems to perform slightly worse than ML, and the tendency that the difference between the algorithms increase with link length is again present.

	M=250		M=500		M=1000	
t_0	f_u^{VML}	f_u^{ML}	f_u^{VML}	f_u^{ML}	f_u^{VML}	f_u^{ML}
0.05	0.09	0.09	0.05	0.04	0.03	0.03
0.075	0.07	0.07	0.04	0.04	0.03	0.02
0.1	0.06	0.06	0.05	0.05	0.03	0.03
0.15	0.07	0.07	0.03	0.02	0.03	0.03
0.2	0.09	0.08	0.06	0.04	0.04	0.03
0.25	0.12	0.10	0.09	0.04	0.07	0.04
0.3	0.11	0.07	0.09	0.07	0.07	0.04

Table 1: Fraction of trees where the correct (generated) tree does not have the lowest free energy (f_u). 500 trees for each link parameter (t_0) and sequence length (M) are tested and standard ML and VML are compared. All tests are performed with $N = 4$ and $K = 4$.

5.3 Primate Sequence Test

VML was also applied to real DNA data, using five homologous primate sequences obtained from the Silver Project[5]. The aligned sequences are from the aromatic L-amino acid decarboxylase (AADC) gene, and 711 nucleotide positions are used from two humans, a chimpanzee, a gorilla and an orangutan.

As there are only five sequences, all possible (15) topologies have been investigated. VML and ML both favor the tree shown in Figure 6 as the most likely tree. The topology is identical to the one proposed along with the data, obtained with a neighbour-joining method. Both ML and VML yielded a free energy per site of $F = 0.04867$, and the link lengths inferred are those given in the figure. Also local search implementations of both the ML and VML methods were tested. A local search implementation starts from a

[5] http://sayer.lab.nig.ac.jp/~silver/

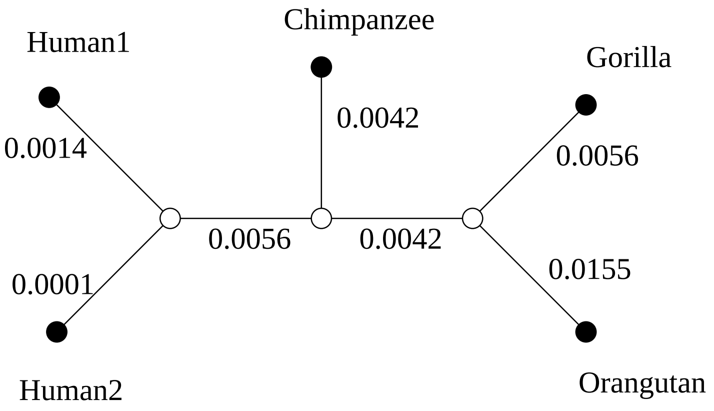

Figure 6: Preferred topology for ML and VML. The inferred link lengths given as expected number of substitutions per site, $h = \frac{K-1}{K}(1-a)$, are the same for the two methods.

random topology, checks the two neighbouring topologies around the shortest internal link, and changes topology if any of the new topologies result in a lower F. This is continued until no improvement can be found. For each choice of an initial topology, the local search variants of ML and VML found the topology shown in figure 6.

6 Summary and Conclusions

We have proposed and explored a novel hybrid approach, VML, combining the variational method with the maximum likelihood principle for the reconstruction of phylogenetic trees based on DNA sequences.

In its most general form, the VML method consists in considering a variationally adjustable parameterized extension of the observed statistics to include also hidden data, such as unknown ancestor sequences. A modified likelihood based on the extended data is then maximized with respect to model parameters as well as the parameters of the data extension. The modified likelihood is shown to approximate the conventional likelihood from below.

For phylogeny reconstruction, such an approach has the advantage that simplifications due to the intrinsic factorization properties of the models can be explored. This enables simpler update equations for the link parameters as compared to standard ML. This is especially apparent with the Jukes-Cantor model, where the link parameters can be directly updated, whereas in standard ML an iterative procedure has to be used.

The method was explored on artificial JC model data, with a simple factorized data extension. Results for cases with reasonably similar sequences were comparable to those of a standard ML approach, and slightly worse for less similar sequences.

Acknowledgements

This work was in part supported by the Swedish Foundation for Strategic Research.

References

Chor, B., M. D. Hendy, B. Holland, and D. Penny (2000). Multiple maxima of likelihood in phylogenetic trees: An analytic approach. *Molecular Biology and Evolution 17*(10), 1529–1541.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution 17*, 368–376.

Felsenstein, J. (1993). *PHYLIP: Phylogenetic Inference Package.* Seattle, WA: University of Washington.

Feynman, R. (1972). *Statistical Mechanics, Frontiers in Physics.* Reading, MA: W. A. Benjamin, Inc.

Friedman, N., M. Ninio, I. Pe'er, and T. Pupko (2002). A structural em algorithm for phylogenetic inference. *Journal of Computational Biology 9*(2), 331–354.

Gu, X. and W. Li (1996). A general additive distance with time-reversability and rate variation among nucleotide sites. *Proceedings of the National Academy of Sciences of the United States of America 93*, 4671–4676.

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

Jukes, T. H. and C. R. Cantor (1969). Evolution of protein molecules. In H. N. Munro (Ed.), *Mammalian Protein Metabolism*, pp. 21–132. New York: Academic.

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

Nei, M. (1996). Phylogenetic analysis in molecular evolutionary genetics. *Annual Review of Genetics 30*, 371–403.

Parisi, G. (1988). *Statistical Field Theory.* Reading, MA: Addison-Wesley Publishing Company.

Swafford, D. L. and G. J. Olsen (1996). Phylogeny reconstruction. In C. M. D. M. Hillis and B. K. Mable (Eds.), *Molecular Systematics.* Sunderland: Sinauer Associates.

Tamura, K. and M. Nei (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Molecular Biology and Evolution 10*, 512–526.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution 39*, 105–111.

Yang, Z. (1997). Paml: A program package for phylogenetic analysis by maximum likelihood (`http://abacus.gene.ucl.ac.uk/software/paml.html`). *CABIOS 13*, 555–556.