Reconstructing phylogeny by Quadratically Approximated Maximum Likelihood

M. D. Woodhams* and M. D. Hendy

Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Private Bag 11-222, Palmerston North, New Zealand

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Summary: Maximum likelihood (ML) methods are popular for phylogenetic inference using genetic sequence data. Suppose we observe aligned homologous DNA sequences \( \Sigma = \{\sigma_0, \sigma_1, \ldots, \sigma_n\} \) from a set \( X \) of \( n+1 \) species, and propose a particular model \( M \) for the evolutionary relationship among \( X \). \( M \) typically will propose a phylogeny \( T \) on \( X \), and a model of nucleotide substitution across the edges of \( T \), described by some selection of parameters \( p(T) \) from a parameter space \( \mathcal{P}(T) \). This specification should be sufficient to calculate the ‘likelihood’

\[
L(\Sigma \mid M, T, p)
\]

of obtaining \( \Sigma \), given the model \( M \) on tree \( T \) with the specified parameters.

The ‘local’ maximization problem of ML is to find

\[
L_{\text{max}}(\Sigma \mid M, T) = \max_{p(T) \in \mathcal{P}(T)} L(\Sigma \mid M, T, p),
\]

for a given phylogeny \( T \). The ‘global’ maximization problem of ML is to find \( T_{\text{max}} \) which maximizes

\[
L_{\text{max}}(\Sigma \mid M, T)
\]

across all trees \( T \).

Typically ‘double optimization’ methods of phylogeny [e.g. Maximum parsimony (MP) and ML] require a search through all potential phylogenies (which can be restricted to all \((2n-3)!经典表达式 in (2) - 1\) binary trees on \( n + 1 \) labelled leaves). The super-exponential (in \( n \)) number of phylogenies limits this complete search to small values of \( n \), so that for many realistically sized problems, ML and MP become heuristic searches. Branch and bound (B&B) searches increase the range of search for MP (Hendy and Penny, 1982), but as yet B&B for ML has been restricted to small cases (Hendy and Holland, 2003).

MP has an \( O(n) \) algorithm (Fitch, 1971) for solving the local optimization, but for most models \( M \) no exact method is available for the local optimization for ML. We introduce here a Quadratic Approximation to Maximum Likelihood (QAML). Although QAML is essentially a heuristic method, preliminary testing has suggested its outputs match ML closely. Further, it is potentially amenable to a B&B search.

The restriction to simple symmetric models of Jukes and Cantor (Jukes and Cantor, 1969) and Kimura’s 2ST and 3ST (Kimura, 1983), enables us to employ the analytic tool of Hadamard conjugation (Hendy et al., 1994). Following (Hendy and Holland, 2003) we develop QAML initially for the simpler 2-state symmetric model introduced by Neymann (Neymann, 1971). (We will comment on its extension to the 4-state models later.)

1.1 Neymann’s model

For Neymann’s model we summarise the relevant information from \( \Sigma \) as a vector \( \hat{s} = (\hat{s}_n)_{\sigma \subseteq [n]} \) (where \( [n] = \{1, 2, \ldots, n\} \)) of split frequencies. Let \( \chi_{ij} \) be the character at the \( j \)-th site of \( \sigma_i \), then the \( j \)-th site induces the sequence split \( \{1/\chi_{ij} = \chi_{0j}\} \cup \{1/\chi_{ij} \neq \chi_{0j}\} \) which we index by the subset \( \alpha = \{1/\chi_{ij} \neq \chi_{0j}\} \subseteq [n] \). \( \hat{s}_\alpha \) is defined to be the relative frequency of the \( \alpha \) site frequency. Thus \( \hat{s} \) is a vector of \( 2^n \) components.
which are conventionally ordered by their indices

\[ \mathbf{s} = (s_0, s_1, s_2, s_3, s_4, s_5, s_6, s_7)^T. \]

\( \mathbf{s} \) is called the ‘observed sequence spectrum’ (Fig. 1).

Likewise the model specification \( M \) on the tree \( T \) is expressed as a vector \( \mathbf{q} \) of splits as follows. Each edge of \( T \) induces a split among the leaves of \( T \), and again we index the edge by the subset not containing 0. For each edge \( e_\alpha \), we specify the parameter \( q_\alpha \), as the ‘edge-length’ being the expected number of substitutions per site across \( e_\alpha \). For example if \( e \) separates 0, 2, 3, 5 from 1, 4, 6 in \( T \), then we index \( e \) as \( e_{\{1,4,6\}} \), and \( q_{\{1,4,6\}} \) (element number \( 41 = 1010011_2 \) in the vector \( \mathbf{q} \)) is the length of edge \( e_{\{1,4,6\}} \).

From \( q_\alpha \) we can calculate

\[ p_\alpha = \frac{1}{2} [1 - \exp(2q_\alpha)], \tag{1} \]

the probability that under a Poisson process, there is an odd number of substitutions, i.e. the probability that the states at the endpoints of \( e_\alpha \) differ. We observe that Equation (1) is readily inverted (provided \( p_\alpha < 0.5 \)) as

\[ q_\alpha = -\frac{1}{2} \ln(1 - 2p_\alpha). \tag{2} \]

The \( q_\alpha \) parameters are collected into the vector \( \mathbf{q} \) of \( 2^n \) components, by setting \( q_\beta = 0 \) for all splits \( \beta \) not in \( T \), together with \( q_\emptyset = -\left( \sum_{e_\alpha \text{ an edge of } T} q_\alpha \right) \cdot \mathbf{q} \) is called an ‘edge-length spectrum’. Its positive components identify the edge splits of \( T \).

### 1.2 Hadamard conjugation

Hadamard conjugation enables us to formulate the probabilities \( s_\alpha [\alpha \subseteq [n]] \) of obtaining each split pattern \( \alpha \) as a function of the edge length spectrum. In particular (Hendy et al., 1994) we can extend Equation (1) to the whole tree \( T \) as

\[ s = H^{-1} \exp(H \mathbf{q}). \tag{3} \]

where the exponential function is applied individually to each component of the vector \( H \mathbf{q} \) and \( H \) is a Hadamard matrix of order \( 2^n \). In particular \( H \) has the following properties which we will use:

- The entries in \( H \) are \( \pm 1 \);
- The leading row and column of \( H \) are all 1s;
- \( H \) is a symmetric orthogonal matrix (hence \( H^{-1} = H^{-T} \));
- If we index the rows and columns of \( H \) by the subsets of \( [n] \) then \( h_{\alpha, \beta} = (-1)^{|\alpha \cap \beta|} \) is -1 if and only if \( \alpha \) and \( \beta \) share an odd number of elements;
- \( \sum_{\beta} h_{\alpha, \beta} = 2^n \delta_{\alpha, \emptyset} \), where \( \delta \) is the Kronecker delta and \( \emptyset \) is the empty set.

For any given tree \( T \), the positive entries of \( \mathbf{s} \) specify the edges of \( T \). Equation (3) enables us to give an analytic formulation of the likelihood,

\[ L(\mathbf{s} | T, \mathbf{q}) = L(\mathbf{s} | \mathbf{q}) = \prod_{\alpha \subseteq [n]} s_\alpha^{q_\alpha}, \tag{4} \]

where \( \mathbf{s} \) is derived from \( \mathbf{q} \) by Equation (3).

### 1.3 Spaces related to Hadamard conjugation

Equation (3) is readily inverted

\[ \mathbf{q} = H^{-1} \ln(H \mathbf{s}), \tag{5} \]

where the natural logarithm is applied to each component of the vector \( H \mathbf{s} \). [N.B. by Equation (3), \( H \mathbf{s} > 0 \).] Hence given \( \mathbf{s} \) we can recover \( \mathbf{q} \) and hence identify \( T \), giving a potential phylogenetic inference tool. Note also the number of independent parameters in \( \mathbf{q} \) is the number of edges of \( T \) (at most \( 2^n - 1 \)).

The conjugation Equation (5) can be dissected into

\[ r = H \mathbf{s}, \quad \rho = \ln(r), \quad \mathbf{q} = H^{-1} \rho. \tag{6} \]

where, like \( \mathbf{q} \) and \( \mathbf{s} \), \( \rho \) and \( \mathbf{r} \) are vectors of \( 2^n \) components which can be indexed by the subsets of \([n] \). As \( H \) is orthogonal these relations are bijective, so each is a function of the (at most \( 2^n - 1 \)) edge lengths \( q_\alpha \), which are constrained to be non-negative.

However an observed sequence spectrum \( \mathbf{s} \) has \( 2^n - 1 \) degrees of freedom, constrained only by \( \sum_{\alpha \subseteq [n]} s_\alpha = 1 \), and by \( s_\alpha > 0 \) for each \( \alpha \subseteq [n] \). We define the ‘s-space’ to be the set of all vectors satisfying these constraints, and note it is the positive quadrant of a hyperplane of dimension \( 2^n - 1 \) embedded in \( \mathbb{R}^{2^n} \). For any \( \mathbf{s} \) in the s-space, we can apply Equation (5), which, provided \( H \mathbf{s} > 0 \), gives

\[ \hat{\mathbf{q}} = H^{-1} \ln(H \hat{\mathbf{s}}), \tag{7} \]

and define the corresponding vectors

\[ \hat{\mathbf{r}} = H \hat{\mathbf{s}}, \quad \hat{\rho} = \ln(\hat{\mathbf{r}}), \quad \hat{\mathbf{q}} = H^{-1} \hat{\rho}. \tag{8} \]

These vectors will lie in subsets of \( \mathbb{R}^{2^n} \), with \( 2^n - 1 \) degrees of freedom, constrained only by

\[ \hat{r}_\emptyset = 1, \quad \hat{\rho}_\emptyset = 0, \quad \sum_{\alpha \subseteq [n]} \hat{q}_\alpha = 0, \tag{9} \]
and some non-negativity conditions. We will refer to these subsets as \( r \)-space, \( p \)-space and \( q \)-space, respectively. (Note these are not Euclidean subspaces.)

For any phylogeny \( T \) on \( n + 1 \) taxa, the set of possible \( q(T) \) vectors will define a subset of \( \mathbb{R}^{2n} \) which we will call the ‘\( q(T) \)-space’ (it will be a bounded subset of a hyperplane of dimension at most \( 2n - 1 \)). The images of all points of the \( q(T) \)-space under the transformations of Equation (6) will be referred to as the ‘\( p(T) \)-space’, the ‘\( r(T) \)-space’ and the ‘\( s(T) \)-space’, respectively. Further, we call the union of these \( s(T) \)-spaces over all phylogenies \( T \) for the same set of \( n + 1 \) taxa, the ‘forest’ space.

### 1.4 Closest Tree

Because any \( \hat{s} \) in \( s \)-space is a finite sample of frequencies, we cannot expect \( \hat{s} \) to belong exactly to \( s(T) \)-space for any tree \( T \). The Closest Tree (CT) algorithm (Hendy, 1989) selects the tree \( T \) for which the Euclidean distance of \( \hat{s} \) to \( s(T) \)-space is minimal. A B&B implementation of CT has been developed and recently implemented in Spectronet, downloadable from http://awcmee.massey.ac.nz/spectronet/.

The logarithm in Equation (7) is estimating evolutionary distances from the number of mutations [it is an extension of Equation (2)]. Other distance functions could be used, such as to account for a distribution of mutation rates across sites, or Tajima’s distance to correct for bias (and inapplicability to negative arguments) of the logarithmic function (Tajima, 1993) which we observe leads to an improvement in CT.

### 2 LIKELIHOOD IN THE HADAMARD SPACES

The likelihood function \( L(\hat{s} \mid T, q) \) of Equation (5) can be extended to a function

\[
L = L(\hat{s} \mid s) = \prod_{a \leq [n], i > 0} s_a^{\hat{s}_a}
\]

over \( s \)-space. In \( s \)-space, \( L \) has a unique maximum at \( s = \hat{s} \). If \( \hat{s} \) is a point in \( s(T) \)-space for some tree \( T \), then \( q \) is in \( q(T) \)-space and \( T = T_{\max} \). However, we must expect that \( \hat{s} \) does not belong to \( s(T) \)-space for any tree \( T \). Hence we seek a tree \( T \) with \( s(T) \)-space close to \( \hat{s} \).

Similarly, when we translate to \( q \)-space, we seek a tree \( T \) with \( q(T) \)-space close to \( \hat{q} \). We can approximate \( L(q \mid \hat{q}) \) in \( q \)-space by a spherically symmetric function, \( L'(q) \), centred at \( \hat{q} \). The point \( q \) closest to \( \hat{q} \) over all \( q(T) \)-spaces maximizes this \( L'(q) \) and is the CT point.

We better approximate the likelihood by allowing \( L'(q) \) to have ellipsoidal contours—i.e. we can approximate \( L(q \mid \hat{q}) \) by a second order polynomial and find the maximum of this \( L' \) in forest-space. We call this the QAML method.

### 2.1 The Quadratic Approximation

A multivariate function \( f(x) \) can be expanded to second order about a stationary point \( x_0 \) by a quadratic form

\[
f(x) \cong (x - x_0)^T A (x - x_0) + f(x_0),
\]

where \( A \) is half the Hessian matrix of the function, evaluated at \( x_0 \):

\[
a_{i,j} = \frac{1}{2} \left. \frac{\partial^2 f}{\partial x_i \partial x_j} \right|_{x=x_0}.
\]

For QAML, we require a quadratic approximation to the likelihood in \( q \)-space, or equivalently in \( p \)-space. Working in \( p \)-space simplifies the calculations. From Equations (4) and (6), the first derivatives of \( L \) are

\[
\frac{\partial L}{\partial p_\alpha} = 2^{-n} L r_\alpha \sum_{\gamma} \hat{q}_\gamma h_{\alpha,\gamma}
\]

for \( \alpha \subseteq [n] \), where here and beyond the summations are over all subsets \( \gamma \subseteq [n] \) with \( \hat{q}_\gamma > 0 \). The second derivatives (Hessian matrix) are

\[
\frac{\partial^2 L}{\partial p_\alpha \partial p_\beta} = 4^{-n} L \left( r_\alpha \sum_{\gamma} \hat{q}_\gamma h_{\alpha,\gamma} \right) \left( r_\beta \sum_{\gamma} \hat{q}_\gamma h_{\beta,\gamma} \right) + 2^{-n} L \left( r_\alpha \sum_{\gamma} \hat{q}_\gamma h_{\alpha,\gamma} \right) \delta_{\alpha,\beta} - 4^{-n} L \left( r_\alpha r_\beta \sum_{\gamma} \hat{q}_\gamma h_{\alpha,\gamma} h_{\beta,\gamma} \right),
\]

for \( \alpha, \beta \subseteq [n] \).

Evaluating Equation (13) at \( p = \hat{p} \):

\[
\frac{1}{L} \frac{\partial^2 L}{\partial p_\alpha \partial p_\beta} \bigg|_{p=\hat{p}} = 2 \delta_{\alpha,\beta} \delta h_{\alpha,\beta} - 4^{-n} \hat{r}_\alpha \hat{r}_\beta \sum_{\gamma} \hat{q}_\gamma h_{\alpha,\gamma} h_{\beta,\gamma}
\]

noting \( \sum_{\beta} h_{\alpha,\beta} = 2^n \delta_{\alpha,\beta} \) and \( r_{\emptyset} = 1 \). Then from Equations (11) and (13) the quadratic form matrix \( A \) has entries

\[
a_{\alpha,\beta} = \frac{1}{2} L(\hat{p}) \left( 2 \delta_{\alpha,\beta} \delta h_{\alpha,\beta} - 4^{-n} \hat{r}_\alpha \hat{r}_\beta \sum_{\gamma} \hat{q}_\gamma h_{\alpha,\gamma} h_{\beta,\gamma} \right).
\]

We now define matrix \( F \) by

\[
f_{\alpha,\beta} = \frac{1}{\sqrt{\hat{s}_\alpha}} \hat{h}_{\alpha,\beta} \hat{r}_\beta, \quad \alpha, \beta \subseteq [n]: \hat{s}_\alpha > 0.
\]

Hence

\[
F = \text{diag} \left( \frac{1}{\sqrt{\hat{s}}} \right) H^{-1} \text{diag} \left( \hat{\theta} \right).
\]

Then the quadratic approximation to \( L \) is

\[
L'(\hat{p}) = L(\hat{p})(\hat{p} - \hat{p})^T (\delta h_{\alpha,\beta} \delta h_{\alpha,\beta} - \frac{1}{2} F^T F) \times (\hat{p} - \hat{p}) + L(\hat{p}).
\]
However, from Equation (9), \( \rho_\emptyset = \hat{\rho}_\emptyset = 0 \), so the \( \delta_{\alpha, \beta} \) term is either zero or multiplied by zero, so has no effect. Furthermore, we are only interested in the location of the maximum, so we can ignore constant terms. Hence we want to maximize

\[
L''(\rho) = -(\rho - \hat{\rho}) F^T F (\rho - \hat{\rho}) = -\| F(\rho - \hat{\rho})\|^2 \\
\]

i.e. we need to minimize \( \| F(\rho - \hat{\rho})\| \) subject to the constraint that \( \rho \in \rho(T) \)-space, for some tree \( T \).

We can modify the \( F \) matrix to accommodate different distance functions, as discussed in Section 1.4. This gives \( F \) as

\[
f_{\alpha, \beta} = \frac{1}{\sqrt{\tilde{s}_\alpha}} \frac{\partial r_{\beta}}{\partial \rho_{\beta}} \bigg|_{\rho = \hat{\rho}}, \quad \alpha, \beta \in [n] : \tilde{s}_\alpha > 0. \tag{20}
\]

The derivation of Equation (20) uses the facts that \( \rho_\emptyset = 0 \) and \( \partial r_{\beta}/\partial \rho_{\beta} = 0 \) if \( \alpha \neq \beta \). [Quadratically approximating the log likelihood (instead of the likelihood) gives the same formation of the Non-Negative Least Squares (NNLS) problem, i.e. we need to minimize \( \| F(\rho - \hat{\rho})\| \) subject to the constraint that \( \rho \in \rho(T) \)-space, for some tree \( T \).]

We can modify the \( \tilde{F} \) matrix by selecting the vectors for \( \tilde{\omega} \) to have the corresponding log likelihood (instead of the likelihood) gives the same formation of the Non-Negative Least Squares (NNLS) problem, i.e. we need to minimize \( \| F(\rho - \hat{\rho})\| \) subject to the constraint that \( \rho \in \rho(T) \)-space, for some tree \( T \).

2.2 Algorithm for finding the QAML tree

We define \( \tilde{q}_\emptyset \) to be the \( q \) vector for the tree where edge \( e_\emptyset \) has unit length and all other edges have zero length. The set \( \{ q_\emptyset \mid e_\emptyset \text{ is an edge in } T \} \) is a basis for \( q(T) \)-space:

\[
q \in q(T) \Rightarrow q = \sum_{e_\emptyset \text{ an edge of } T} q_\emptyset \tilde{q}_\emptyset, \quad \text{all } q_\emptyset \geq 0. \tag{21}
\]

These \( q(T) \)-space basis vectors can be transformed into \( \rho' \)-space:

\[
\rho' = F \rho = F \tilde{H} q \Rightarrow \rho' = \sum_{e_\emptyset \text{ an edge of } T} q_\emptyset \tilde{\rho}_\emptyset, \quad \text{all } q_\emptyset \geq 0 \tag{22}
\]

where \( \tilde{\rho}_\emptyset = F \tilde{H} \tilde{q}_\emptyset \). Now define \( x_T \) to be the vector of the \( 2n - 1 \) \( q_\emptyset \) values for \( e_\emptyset \) in \( T \), and matrix \( B_T \) to have the corresponding \( \tilde{\rho}_\emptyset \) basis vectors as its columns, then Equation (22) becomes

\[
\rho' = B_T x_T \tag{23}
\]

and the quantity we seek to minimize is

\[
\| F(\rho - \hat{\rho})\| = \| \rho' - \hat{\rho}' \| = \| B_T x_T - \hat{\rho}' \| \tag{24}
\]

(\( \hat{\rho}' = F \hat{\rho} \), subject to the constraint \( x_T \geq 0 \). This is in the form of the Non-Negative Least Squares (NNLS) problem, for which an algorithmic solution is in Lawson and Hanson (1995).

We could at this point produce an exhaustive search algorithm that applies NNLS for each tree. It is more efficient to apply QR decomposition to find the unconstrained minimum, a lower bound on the constrained (NNLS) minimum. NNLS need only be applied to a small subset of trees, as described in the following algorithm:

1. Calculate \( \hat{\rho}' = F \hat{\rho} \).
2. Calculate and cache all the \( \tilde{\rho}_\emptyset = F \tilde{H} \tilde{q}_\emptyset \) vectors for all \( \alpha \subseteq [n] \):
3. For each possible tree \( T \):
   (a) Create the \( B_T \) matrix by selecting the vectors for \( T \) from those calculated in stage 2
   (b) Use QR decomposition to solve the unconstrained problem
      \[
      R(T) = \min_{x_T} \| B_T x_T - \hat{\rho}' \|
      \]
   (4) Apply NNLS to solve the constrained problem \( Q(T) = \min_{x_T \geq 0} \| B_T x_T - \hat{\rho}' \| \) for the trees with the lowest \( R(T) \) scores until all remaining trees have a \( R(T) \) score higher than the best \( Q(T) \) score found.
4. The QAML tree is given by the tree topology \( T \) that gave the smallest \( Q(T) \), with edge lengths given by \( x_T \) for that tree.

2.3 Computational complexity

The major contributions to the computational complexity of the algorithm are

1. Multiplications by \( F \) in steps 1 and 2 are \( O(n 2^n) \).
2. \( 2^n - 1 \) vectors are calculated in step 2.
3. There are \( (2n - 3)! \) trees considered in step 3.
4. Each QR decomposition in step 3b is \( O(mn^2) \) where \( m \) is the number of non-zero elements of \( \tilde{\omega} \).

We are investigating techniques to reduce the computational complexity of this algorithm for finding the QAML tree, including the possibility of a B&B algorithm similar to that for CT.

3 EXTENSIONS OF THE QAML METHOD

3.1 Four state data

The Hadamard Conjugation method is extended to the Kumura 3ST model by Hendy, Penny and Steel (Hendy et al., 1994). In this extension, each of the Hadamard vectors \( (s, r, \rho, q) \) are of length \( 4^n \). Each edge of \( T \) has three entries in \( q \), for transitions and type I and II transversions. QAML can be applied to these models with an appropriate choice of basis vectors. If we allow transition/transversion ratios to vary from edge to edge, we use three basis vectors per edge, each with a single non-zero element in \( q \)-space. For a Kimura 3ST model,
we use a single basis vector per edge, with three non-zero elements in ratios equal to the transition/transversion ratios we have chosen.

The Kimura 2ST model (type I and II transversion rates are equal) and Jukes Cantor model (transversion and transition rates are all equal) are special cases of Kimura 3ST.

Of the $4^n$ elements of $q$, all but $3 \times 2^n - 2$ are zero in all models. Waddell and Hendy have shown how to calculate just the relevant elements of $\tilde{q}$ in $O(n 2^n)$ time instead of $O(n 4^n)$, as would be the case if the full Hadamard conjugation were applied to the extended vectors (Waddell and Hendy, 1997).

3.2 Molecular clock

QAML can be applied with a molecular clock, by choice of basis vectors plus additional inequality constraints. An algorithm to achieve this is as follows. Starting with a rooted tree topology, add $n - 1$ labels to edges of the binary rooted tree $T$ according to the following rules:

1. Every internal node has one labelled descendent edge and one unlabelled
2. If exactly one descendent edge of a node is terminal (has an external node/taxon at the lower end), label the non-terminal edge.

(See Fig. 2 for an example.) Then each edge of the unrooted tree can be expressed as a linear combination of the labelled edges, e.g.

$$q[1] = a_1 + 0 a_2 + 0 a_3 + 0 a_4,$$

$$q[2] = a_1 + 0 a_2 + 0 a_3 + 0 a_4,$$

$$q[1,2] = a_1 + 2 a_2 - 1 a_3 - 1 a_4,$$

$$q[3] = 0 a_1 + 0 a_2 + 1 a_3 + 1 a_4,$$

$$q[1,2,3] = 0 a_1 + 0 a_2 + 0 a_3 + 1 a_4,$$

$$q[4] = 0 a_1 + 0 a_2 + 1 a_3 + 0 a_4,$$

$$q[1,2,3,4] = 0 a_1 + 0 a_2 + 1 a_3 + 0 a_4.$$  \hfill (25)

The columns of Equation (25) give us the required basis vectors:

$$\tilde{a}_1 = \tilde{q}[1] + \tilde{q}[2] + \tilde{q}[1,2],$$

$$\tilde{a}_2 = 2 \tilde{q}[1,2],$$

$$\tilde{a}_3 = -\tilde{q}[1,2] + \tilde{q}[3] + \tilde{q}[4] + \tilde{q}[1,2,3,4],$$

$$\tilde{a}_4 = -\tilde{q}[1,2] + \tilde{q}[3] + \tilde{q}[1,2,3].$$ \hfill (26)

So we construct a $B$ matrix from these vectors and find $x$ to minimize $\|Bx - \hat{\rho}\|$. As before, we have the constraint that $x \geq 0$, but to ensure that $q_{[1,2]} \geq 0$ we have an additional constraint that $x_1 + 2x_2 - x_3 - x_4 \geq 0$. (Every unlabelled internal edge will yield such a constraint. Rule 2 in the labelling algorithm exists only to minimize the number of such constraints.) This class of problem is known as Least Squares with Inequalities (LSI), the solution to which is closely related to NNLS (Lawson and Hanson, 1995).

4 MONTE CARLO TESTING

4.1 Phylogeny selection

Monte Carlo data was generated for the tree in Figure 3, for sequence lengths of 100, 200, 300, 500, 700 and 1000 bp. Data were produced under the two state Cavender/Farris model, and the four state Kimura 2ST model, with transition : transversion ratio of 2. Each model/sequence length combination had 5000 simulations.

CT, QAML and ML were applied to each simulation [CT and QAML were done using both the log distance and Tajima distance (Tajima, 1993)]. The four state data was analysed assuming fixed transition/transversion ratio of 2.

Figure 4 plots the relative error rates (how often the wrong tree topology was chosen, normalized so that ML always has an error rate of 1). Both QAML and ML were 100% accurate for 700 or 1000 four-state bases (not plotted). QAML outperforms CT, and ML marginally outperforms QAML. Tajima distances consistently improve the accuracy of CT, and give a modest improvement to QAML.

In our test implementation, QAML took about three times as long to analyse as CT. There was no measurable time penalty for using Tajima distances. ML analyses were done with PAUP* (Swofford, 2003). As our test implementation is unoptimized and in an interpreted language (Mathematica), we cannot sensibly compare run times between ML and QAML with the current implementation.

4.2 Edge length estimation

We have measured the accuracy of the edge length estimates of the various methods on random nearly clock-like trees. The accuracy is measured by the statistic $D$ (which we call the
deviation) given by

\[ D_{\text{method}} = \frac{\sum (q_\alpha - \hat{q}_{\text{method}, \alpha})^2}{\sum q_\alpha^2}, \] (27)

where \( \hat{q}_{\text{method}, \alpha} \) is the edge length estimate for split \( \alpha \) for the given method (CT, QAML or ML).

To generate a random almost clock-like \( n \)-taxon tree, we first generate a clock-like \( n + 2 \) taxon tree where speciation events occur via a Poisson process at constant rate on all branches. We prune 2 randomly chosen taxa to reduce the tree to \( n \) taxa. Each edge length is then multiplied by a random factor, drawn from a Gaussian distribution with mean 1 and SD 0.1. The speciation rate for five taxon trees was selected so that the sum lengths of the branches had a median value of 0.8, the same as the tree in Figure 3. (The speciation rate that achieved this was once per \( 1.44 \times 10^{-4} \) mutations per base.) The speciation rate for 10 taxon trees was slightly lower, at one speciation per \( 1.14 \times 10^{-4} \) mutations per base.

For each of 5 and 10 taxa, 2000 trees were generated. For each tree, two state sequences were randomly generated with lengths 100, 200, 300, 500, 700 and 1000 bases. Each set of sequences were analysed, for the known correct phylogeny, by CT and QAML (using log and Tajima distances) and ML. The deviations [Equation (27)] were calculated, and the median deviation for each method/sequence length was found. These values are plotted in Figure 5, for CT and QAML relative to ML. (The deviations for ML are inversely proportional to sequence length.)

As in Figure 4, QAML outperforms CT. Apart from that, the results are varied. QAML performs similarly to ML for five taxa, but much worse for ten taxa. Tajima distances give similar or superior results for five taxa, but worse results for ten taxa.

5 PRIMATE PSEUDogene ANALYSIS

Haemoglobin \( \psi \)-pseudogene sequences of length 6771 were analysed for the five primate species human, chimpanzee, gorilla, orangutan and rhesus monkey (Miyamoto et al., 1988; Hendy et al., 1994). Of the sites with exactly two distinct bases, 460 are transitions, 107 type I transversions and 86 type II transversions. The data were analysed using the Tajima distance function and a Kimura 2 substitution type model with transition:transversion ratio of 2.375 : 1, derived from the counts of sites with two distinct bases. Figure 6a shows the edge lengths for CT, QAML and ML without any molecular clock constraint. Figure 6b shows the QAML and ML edge lengths with a molecular clock constraint imposed. If we take QAML
Fig. 6. The tree derived from haemoglobin $\psi$-pseudogene DNA sequences under a Kimura 2 substitution type model. Edge lengths are in, mutations, per thousand bases. (a) Unconstrained edge lengths. Italic, CT, plain, QAML; **bold**, ML. (b) Edge lengths with molecular clock constraint. Plain, QAML; **bold**, ML.

the root of this tree to be 23.3 Mya (Kumar and Hedges, 1998), then the internal nodes are 3.8, 4.3 and 9.5 Mya (QAML) or 4.5, 5.2, and 10.0 Mya (ML).

6 DISCUSSION

The QAML method is based on CT, but has accuracy approaching that of ML. We have conducted preliminary Monte Carlo investigations into its accuracy, and demonstrated its use on real data, both with and without a molecular clock.

If QAML can be implemented efficiently (largely reliant on incorporating an efficient B&B algorithm) it could become a useful tool to speed up ML analysis: the best few QAML trees can be used as starting points for the ML optimization. A fundamental limitation of QAML is its reliance on the Kimura 3ST model (and submodels).

ACKNOWLEDGEMENTS

The authors wish to acknowledge the assistance of Bettina Greese and Anke Schulz, exchange students from the University of Greifswald, Germany, who extracted, aligned and analysed the sequence data for Figure 6, in their semester project at Massey University, 2003.

REFERENCES


