Maximum likelihood estimation of phylogenetic tree and substitution rates via generalized neighbor-joining and the EM algorithm

Asger Hobolth
Bioinformatics Research Center, 1500 Partners II, 840 Main Campus Drive, North Carolina State University, Raleigh, NC 27606, USA, asger@statgen.ncsu.edu
Ruriko Yoshida*
Department of Mathematics, Duke University Box 90320, Durham, NC 27708-0320, USA, ruriko@math.duke.edu

Keywords: EM algorithm, neighbor joining, phylogenetic reconstruction, subtree weights

∗ Corresponding author.

Abstract

A central task in the study of molecular sequence data from present-day species is the reconstruction of the ancestral relationships. The most established approach to tree reconstruction is the maximum likelihood (ML) method. In this method, evolution is described in terms of a discrete-state continuous-time Markov process on a phylogenetic tree. The substitution rate matrix, that determines the Markov process, can be estimated using the expectation maximization (EM) algorithm. Unfortunately, an exhaustive search for the ML phylogenetic tree is computationally prohibitive for large data sets. In such situations, the neighbor-joining (NJ) method is frequently used because of its computational speed. The NJ method reconstructs trees by clustering neighboring sequences recursively, based on pairwise comparisons between the sequences. The NJ method can be generalized such that reconstruction is based on comparisons of subtrees rather than pairwise distances. In this paper, we present an algorithm for simultaneous substitution rate estimation and phylogenetic tree reconstruction. The algorithm iterates between the EM algorithm for estimating substitution rates and the generalized NJ method for tree reconstruction. Preliminary results of the approach are encouraging.

1. Introduction

Current efforts to reconstruct the tree of life for different organisms demand the inference of phylogenies from thousands of DNA sequences. The first author is at a university where the tree of life for flies is investigated (http://www.inhs.uiuc.edu/cee/FLYTREE/) and the second author is

at a university where the tree of life for fungi is considered (http://ocid.nacse.org/research/aftol/). For such large data sets the tree space is enormous and identification of the optimal tree is a major challenge.

The evolution of homologous DNA sequences can be described by continuous time Markov chains on a phylogenetic tree [8]. A continuous time Markov chain is characterized by a substitution rate matrix, and the phylogenetic tree summarizes the relationships between the species in terms of edge lengths (times since divergence) and common ancestors. The DNA sequences are only observed in the leaves, and information on the phylogenetic tree, substitution events (time and type) and edge lengths is missing. The transition matrix $P(t)$ for a continuous time Markov process can be written as $\exp(Qt)$, where Q is a parameterized substitution rate matrix. In order to estimate the rate parameters and edge lengths for a fixed tree and based on the observed data, one can use the expectation maximization (EM) algorithm [10]. The updating step of the EM algorithm can be written explicitly in terms of eigenvalues and eigenvectors of Q. The continuous time Markov process gives rise to a distance measure between any sets of sequences. Pairwise distances can be used, together with the neighbor joining (NJ) algorithm, to reconstruct the phylogenetic tree that relates the sequences. The generalized neighbor joining (GNJ) algorithm [11] improves the NJ algorithm by using distance measures based on subtrees rather than pairwise distances.

In this paper, we describe an algorithm that simultaneously estimates the substitution rate matrix and reconstructs the phylogenetic tree. The algorithm iterates between the EM algorithm for rate matrix estimation and the GNJ algorithm for phylogenetic tree reconstruction. We are in the process of implementing the algorithm, and preliminary results are encouraging.

2. The EM algorithm

2.1. Two sequences

Consider the somewhat unreasonable situation where we have access to the *complete observation* (the *hidden model* or the *complete data model*) of the evolution of a single site. The state of the process at time t is denoted $x(t)$, and we have observed the process from time $t = 0$ to time $t = T$. In this paper the size α of the state space is 4 corresponding to the four nucleotides $\Sigma =$`{A,G,C,T}`, but our method can also be applied to protein-coding sequences where the state space is of size 61 because there are 61 sense codons. Continuous-time Markov processes are described in e.g. [9].

We now derive the likelihood for a complete observation of a continuous-time Markov process with substitution rate matrix Q. Suppose the evolution of a single site has been completely observed from time $t = 0$ to time $t = T$ and is given as in Fig. 1. We model the evolution in terms of a continuous-time Markov process with substitution rate matrix Q. Recall that a rate matrix has non-negative off-diagonal entries and each row sums to zero. Let $Q(a, b)$ be the entry of Q in the ath

column and the bth row. The waiting time in a state a is exponentially distributed with parameter $-Q(a,a)$ and the probability of substituting a with b is proportional to $Q(a,b)$. Thus, the likelihood for the complete observation in Fig. 1 is given by

$$\begin{aligned} L(Q) &= Q(1,1)e^{Q(1,1)t_1}\frac{Q(1,3)}{Q(1,1)}Q(3,3)e^{Q(3,3)(t_2-t_1)}\frac{Q(3,1)}{Q(3,3)}Q(1,1)e^{Q(1,1)(t_3-t_2)}\frac{Q(1,2)}{Q(1,1)}e^{Q(2,2)(T-t_2)} \\ &= e^{Q(1,1)(t_1+(t_3-t_2))+Q(2,2)(T-t_2)+Q(3,3)(t_2-t_1)}Q(1,2)Q(1,3)Q(3,2), \end{aligned}$$

where indices 1,2,3,4 corresponds to `A`,`G`,`C`,`T`.

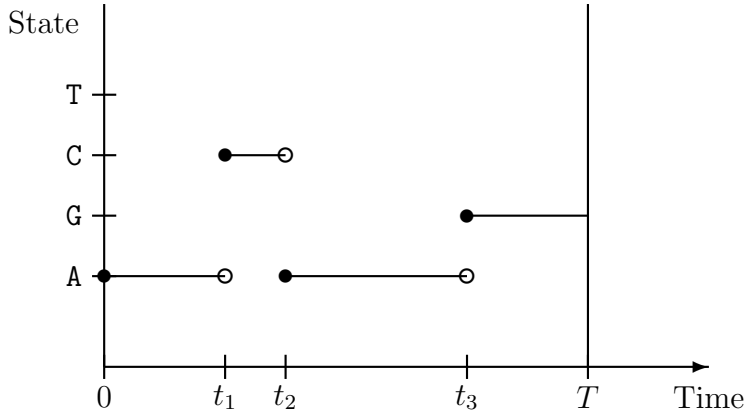

Fig. 1. Complete observation of the evolution of a single site in a DNA sequence.

More generally, if the rate matrix is parametrized by θ, $Q = Q_\theta$, and with $x = \{x(t) : 0 \le t \le T\}$, the maximum likelihood estimation problem for a complete observation is:

$$\text{maximize} \quad L(\theta; x) = \left[\prod_{a\in\Sigma}\prod_{b\neq a} f_{ab}(\theta)^{N(a,b)}\right]\left[\prod_{a\in\Sigma} f'_a(\theta)^{T(a)}\right] \quad \text{subject to} \quad \theta \in \Theta. \tag{1}$$

Here $f_{ab}(\theta)$ and $f'_a(\theta)$ are functions from $\mathbb{R}^d \to \mathbb{R}_+ = \{x \in \mathbb{R} : x > 0\}$ such that $f_{ab}(\theta) = Q_\theta(a,a)$, $f'_a(\theta) = \exp(Q_\theta(a,a))$ for $\theta \in \Theta \subset \mathbb{R}^d$, $T(a)$ is the total time spent in state a, and $N(a,b)$ is the number of substitutions of a with b. Thus, the log-likelihood for a complete observation becomes

$$\log L(\theta; x) = \sum_{a\in\Sigma} T(a)Q_\theta(a,a) + \sum_{a\in\Sigma}\sum_{b\neq a} N(a,b)\log Q_\theta(a,b). \tag{2}$$

Note that the complete log-likelihood is linear in the total time spent in a state and the number of substitutions between states. The complete log-likelihood function is analytically tractable. Indeed, as argued in [10], in most Markov processes of sequence evolution the maximum likelihood estimates can be found analytically from a complete observation.

Consider for example the general time reversible (GTR) model. Let $\pi_a,\ a \in \Sigma,\ \sum_a \pi_a = 1$, denote the stationary distribution of the Markov chain. This distribution can be estimated from the nucleotide frequencies in a single sequence. The GTR model has substitution rate matrix (e.g., [20])

$$Q_\theta = \begin{bmatrix} \cdot & \theta_{12}\pi_2 & \theta_{13}\pi_3 & \theta_{14}\pi_4 \\ \theta_{12}\pi_1 & \cdot & \theta_{23}\pi_3 & \theta_{24}\pi_4 \\ \theta_{13}\pi_1 & \theta_{23}\pi_2 & \cdot & \theta_{34}\pi_4 \\ \theta_{14}\pi_1 & \theta_{24}\pi_2 & \theta_{34}\pi_3 & \cdot \end{bmatrix} \tag{3}$$

where the diagonal elements are such that each row sums to zero and the 6 unknown parameters are $\theta = (\theta_{12}, \theta_{13}, \theta_{14}, \theta_{23}, \theta_{24}, \theta_{34})$. A simple calculation shows that the complete log-likelihood (2) is maximized for

$$\theta^*_{ab} = \frac{N(a,b) + N(b,a)}{\pi_b T(a) + \pi_a T(b)}, \quad a < b.$$

The problem is, however, that the DNA sequences are only observed in the leaves, whereas information on substitution events (time and type) and edge lengths is missing. For two sequences we only observe the beginning state $x(0)$ and the final state $x(T)$ of the process. We have a *missing data problem* in the sense that we only have access to part of the data.

The Expectation Maximization (EM) algorithm [3] is a broadly applicable approach for missing data problems. The algorithm is an iterative procedure with two steps in each iteration, the Expectation Step (E-step) and the Maximization Step (M-step). The algorithm approaches the problem of solving the actually observed log-likelihood indirectly by proceeding iteratively in terms of the complete log-likelihood. As the complete log-likelihood is unobservable, it is replaced by its conditional expectation given the observed data $y = \{x(0), x(T)\}$. In the E-step of the $(k+1)$'th iteration, the function $G(\theta; \theta^k) = \mathrm{E}_{\theta^k}[\log L(\theta; x)|y]$ is calculated, and in the M-step a new parameter value θ^{k+1} is obtained as the value of θ that maximizes $G(\theta; \theta^k)$. For the GTR model (3) the M-step becomes

$$\theta^*_{ab} = \frac{\mathrm{E}_{\theta^k}[N(a,b)|x(0),x(T)] + \mathrm{E}_{\theta^k}[N(b,a)|x(0),x(T)]}{\pi_b \mathrm{E}_{\theta^k}[T(a)|x(0),x(T)] + \pi_a \mathrm{E}_{\theta^k}[T(b)|x(0),x(T)]}, \quad a < b.$$

The algorithm converges to a local maximum likelihood estimate of the observed data.

Since the complete log-likelihood is linear in the time spent in a state and the number of transitions between states, and expectation is a linear operator, all we need in the E-step is to calculate conditional expectations of these two quantities given the observed data. The conditional expectations are provided in the following theorem.

Theorem 1 (Conditional expectations [10])**.** *Consider a continuous time Markov process* $\{x(t) : 0 \leq t \leq T\}$ *on a finite state space with rate matrix* Q*. Denote the transition matrix* $P(t) = \exp(Qt)$*. We have the following conditional expectations:*

- *Time spent in state a*

$$\mathrm{E}[T(a)|x(0)=i,x(T)=j]=\int_0^T P_{ia}(t)P_{aj}(T-t)dt/P_{ij}(T).$$

- *Number of transitions between states a and b*

$$\mathrm{E}[N(a,b)|x(0)=i,x(T)=j]=Q(a,b)\int_0^T P_{ia}(t)P_{bj}(T-t)dt/P_{ij}(T).$$

The transition probability matrix $P(t)=\exp(Qt)$ with entries $P_{ab}(T)=P(X(T)=b|X(0)=a)$ is calculated using an eigenvalue decomposition of Q. Let U be the orthogonal matrix with eigenvalues as columns and D_λ the diagonal matrix of corresponding eigenvectors such that $Q=UD_\lambda U^{-1}$ and $P(T)=e^{QT}=Ue^{TD_\lambda}U^{-1}$. In the most general case some of the eigenvalues and eigenvectors are complex. Conditional expectations are now found from

$$\int_0^T P_{ab}(t)P_{cd}(T-t)dt=\sum_i U_{ai}U^{-1}_{ib}\sum_j U_{cj}U^{-1}_{jd}J_{ij},\ \text{where } J_{ij}=e^{T\lambda_j}\int_0^T e^{t(\lambda_i-\lambda_j)}dt.$$

Even when the eigenvalues are complex, the J_{ij} integrals are easy to evaluate.

In the case of multiple sites that are identically distributed and evolve independently we can summarize the data in a frequency table $\nu(i,j)$ that summarizes the number of sites with $x(0)=i$ and $x(T)=j$. In this case the complete log-likelihood conditional on the observed data becomes

$$G(\theta;\theta^k)=\sum_i\sum_j \nu(i,j)\mathrm{E}_{\theta^k}[\log L(\theta;x)|x(0)=i,x(T)=j].$$

The linearity in the time spent in a state and the number of jumps between states is maintained and the EM-algorithm remains simple. Indeed, the conditional expectations are given by Theorem 1 and the M-step for the GTR model becomes

$$\theta^*_{ab}=\frac{\sum_i\sum_j \nu(i,j)(\mathrm{E}_{\theta^k}[N(a,b)|x(0)=i,x(T)=j]+\mathrm{E}_{\theta^k}[N(b,a)|x(0)=i,x(T)=j])}{\sum_i\sum_j \nu(i,j)(\pi_b\mathrm{E}_{\theta^k}[T(a)|x(0)=i,x(T)=j]+\pi_a\mathrm{E}_{\theta^k}[T(b)|x(0)=i,x(T)=j])}.$$

Note that because we assume $P(t)=\exp(Qt)$, t and Q are confounded. Indeed, $Qt=(2Q)/(t/2)$, which means that twice the rate at half the time has the same results. To avoid confounding, the rate matrix is calibrated such that time corresponds to expected changes per site. This means that $\sum_i\sum_{j\neq i}\pi_i Q_{ij}=1$.

2.2. Multiple sequences

Now consider the case of multiple sequences related by an unrooted phylogenetic tree with n leaves (terminal nodes), n terminal edges, $n-2$ internal nodes and $n-3$ internal edges. The single site complete log-likelihood becomes

$$\log L(\theta; x) = \sum_{i=1}^{2n-3} \Big(\sum_{a\in\Sigma} T^i(a) Q^i(a,a) + \sum_{a\in\Sigma} \sum_{b\neq a} N^i(a,b) \log Q^i(a,b) \Big)$$

where $T^i(a)$ is the total time spent in state a on edge i and $N^i(a,b)$ is the number of transitions from a to b on edge i. Letting $y = (y^1, \ldots, y^n)$ be the observed data at the leaves and letting $a(i), d(i)$ be the ancestral and descendant values at the two ends of the edge with descendant node i we get from the Markov property

$$G(\theta; \theta^k) = \sum_{i=1}^{2n-3} \sum_{a(i),d(i)} \mathrm{E}_{\theta^k} \Big[\sum_{a\in\Sigma} T^i(a) Q^i(a,a) + \sum_{a\in\Sigma} \sum_{b\neq a} N^i(a,b) \log Q^i(a,b) | a(i), d(i) \Big] P_{\theta^k}(a(i), d(i)|y).$$

Here $P_{\theta^k}(a(i), d(i)|y)$ is the probability of observing the ancestral value $a(i)$ and descendant value $d(i)$ at the edge with descendant node i given the data y. These probabilities are calculated using Felsenstein's peeling algorithm [6]. In the E-step we therefore need to calculate conditional mean values on each edge. Conditioning on $a(i)$ and $d(i)$, the conditional mean values are determined by Theorem 1

In this paper we consider the case where the rate matrix is the same on all edges, but the length varies between edges. For example, the rate matrix Q could be the GTR rate matrix (3) calibrated.The rate matrix on each edge i is then given by $Q^i = w_i Q, \;\; i = 1, \ldots, 2n-3$, where w_i is the length of the edge with descendant node i. The M-step is slightly more complicated for multiple sequences compared to pairwise sequences, but can be carried out as described in [10].

3. Generalized neighbor-joining

We will consider binary trees, meaning that all terminal nodes are of degree one and all internal nodes are of degree three. Given a tree, its edge lengths are said to be additive if the distance between any pair of leaves is the sum of the lengths of the edges on the path connecting them. Homogeneous finite-state continuous time Markov models satisfy

$$\sum_k P_{ik}(t_1) P_{kj}(t_2) = P_{ij}(t_1 + t_2). \tag{4}$$

Define the maximum likelihood distance [5] between a pair of leaves $\{i, j\}$ by

$$d(i,j) = \arg\max_t \Big\{ \prod_{s=1}^{N} P_{y^i(s), y^j(s)}(t) \Big\}$$

where $y^i = (y^i(1), \ldots, y^i(N))$ is the DNA sequence of length N observed at leaf i. Suppose the leaves i and j are connected by node k. Using (4) and consistency of maximum likelihood we get $d(i,j) \approx t_1 + t_2$. Extending this argument to a general phylogenetic tree we see that maximum likelihood distances based on continuous time Markov chains between leaf sequences should be close to additive if there is enough data to obtain reliable estimates of the pairwise distances.

3.1. Saitou-Nei neighbor-joining method

Given a binary tree Γ with additive edge lengths, we can reconstruct it from the pairwise distances $d(i,j)$ of its leaves $\{i, j\}$ using the neighbor-joining method of Saitou and Nei (1987). Firstly, pick a cherry $\{i, j\}$ in the tree, i.e. leaves that have the same parent node l. Secondly, remove the leaves i and j from the set of leaves and add l, defining its distance to any other leaf k by

$$d(l,k) = \frac{1}{2}(d(i,k) + d(j,k) - d(i,j))$$

and edge lengths w_i and w_j of edges with descendant nodes i and j, respectively, by

$$w_i = \frac{1}{2}\left[d(i,j) + \frac{1}{n-2}\sum_{k\neq i, k\neq j}(d(i,k) - d(j,k))\right] \quad (5)$$

$$w_j = d(i,j) - w_i. \quad (6)$$

Repeat the procedure until the number of leaves n is 3. The main problem is picking the cherry, and a solution was suggested in Saitou and Nei [17] and modified by Studier and Keppler [18].

Theorem 2 (Cherry picking criterion [17, 18])**.** *Let d be an additive tree metric and define the $n \times n$-matrix B with entries*

$$B(i,j) = (n-2)d(i,j) - \sum_{k\neq i} d(i,k) - \sum_{k\neq j} d(j,k).$$

Then the pair of leaves $\{i^*, j^*\}$ that minimizes B is a cherry in the tree Γ.

3.2. The generalized neighbor-joining method

It is natural to ask if we can generalize the neighbor-joining method based on pairwise distances to distances based on subtrees. Here, a subtree is a set of leaves and a subtree distance is the sum of the lengths of the edges on the path connecting the subtree leaves. Distances based on subtrees are expected to be more accurately determined than pairwise distances because of the following reasons: (1) they are calculated from more data, and (2)we include more conditions to the calculation. (When

we calculate pairwise distances via MLEs, we assume that each path containing a pair is independent, while they are not independent because each path share some branches with the other.) Pachter and Speyer [13] show that binary trees with additive edge lengths are indeed determined from distances based on subtrees. Let Ω be set of leaves in the tree Γ and let $\binom{\Omega}{m}$ be the set of all m-subsets of Ω. Then we denote $D(R)$ for $R \in \binom{\Omega}{m}$ the subtree distance of the subtree containing $R \in \binom{\Omega}{m}$.

Theorem 3 (Subtree distance condition [13]). *Let Γ be binary tree with additive edge lengths. Let m, $2 \leq m \leq (n+1)/2$, be the size of each subtree. The tree Γ is uniquely determined from the set $\{D(R) : R \in \binom{\Omega}{m}\}$.*

Pachter and Speyer do not describe how to reconstruct the tree and calculate the edge lengths from the subtree distances. The tree reconstruction can be found in [11] and is a two-stage procedure.

Theorem 4 (Equivalent topologies [11]). *Suppose that $\{i, j\}$ is a pair of leaves in the set of leaves Ω with $n = |\Omega|$ and suppose that $2 \leq m \leq n-2$. Define the pairwise distances*

$$\tilde{d}(i,j) = \sum_{\Lambda \in \binom{\Omega \setminus \{i,j\}}{m-2}} D(\{i,j,\Lambda\}). \tag{7}$$

The tree $\tilde{\Gamma}$ based on the additive tree metric $\tilde{d}$ has the same topology as the tree Γ based on the additive tree metric D.

We can thus find the topology of the tree Γ by using the cherry picking criterion in Theorem 2 on the pairwise distances $\tilde{d}(i,j)$. Using equations (5) and (6) we can also find the edge lengths $\tilde{w}_i$ for any edge $\tilde{e}_i$ in the tree $\tilde{\Gamma}$. In [11] the map between the edge lengths in $\tilde{\Gamma}$ and the edge lengths in Γ is also described. The map between the edge lengths in the two metrics is linear and one-to-one and given as follows.

Theorem 5 (Map between edge lengths [11]). *Let Γ, $\tilde{\Gamma}$, m and n as in Theorem 4. and let $L_j(e)$ denote the set of leaves in the component of $\Gamma - e$ (or equivalently $\tilde{\Gamma} - e$) that contains leaf j after removing an edge e.*

1. *Let e_i, $i = n+1, \ldots, 2n-3$, denote the internal edges of Γ with lengths w_i and let $\tilde{e}_i$ be the corresponding internal edges of $\tilde{\Gamma}$ with lengths $\tilde{w}_i$. Then*

$$w_i = \frac{2\tilde{w}_i}{\binom{|L_a(e_i)|-2}{m-2} + \binom{|L_c(e_i)|-2}{m-2}}, \quad i = n+1, \ldots, 2n-3,$$

where a and c are the leaves on opposite sides of the edge e_i.

2. *Denote the terminal edges of* Γ *by* $e_1, \ldots, e_n$ *with corresponding edges* $\tilde{e}_1, \ldots, \tilde{e}_n$ *in* $\tilde{\Gamma}$. *Let*

$$C_i = \sum_{j=n+1}^{2n-3} \left(\binom{n-2}{m-2} - \binom{|L_i(e_j)|-2}{m-2} \right) w_j$$

Then

$$\begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = A \begin{pmatrix} 2\tilde{w}_1 - C_1 \\ \vdots \\ 2\tilde{w}_n - C_n \end{pmatrix}, \text{ where } A = \frac{1}{2\binom{n-3}{m-2}} \left(I - \frac{m-2}{(m-1)(n-2)} J \right).$$

Here, I *is the identity matrix and* J *is the matrix consisting entirely of* 1*s.*

The running time for computing the subtree distances is $O(Ln^m)$ where L is the length of the alignment and the computation of the distance matrix $\tilde{d}$ is $O(n^m)$ (both steps are trivially parallelizable). The subsequent neighbor-joining is $O(n^3)$ and edge weight reconstruction is $O(n^2)$. It is interesting to note that for fixed L the running time of the algorithm is $O(n^3)$ for both $m = 2$ and $m = 3$.

4. The EMGNJ algorithm

Suppose we wish to estimate the GTR model. The EMGNJ algorithm consists of two main parts, reconstructing a tree via the GNJ method and improving GTR rates via the EM algorithm. We iterate these steps until it converges.

Algorithm 6 (The EMGNJ algorithm)**.** *Suppose we have n DNA sequences and an integer* $2 \leq m \leq n-2$.

Input*: n DNA sequences and an integer* $2 \leq m \leq n-2$.
Output*: The GTR rates and a phylogenetic tree.*

1. *Estimate stationary distribution from empirical frequencies.*
2. *Reconstruct tree using MJOIN under the Jukes-Cantor (JC69) model.*
3. *Estimate GTR substitution rates and edge lengths from current tree.*
4. *Reconstruct tree using MJOIN and current GTR rates.*
5. *If likelihood is not improved return current tree and GTR rates; otherwise go to 3.*

In other words, we provide starting values of the algorithm in Step 1 and Step 2 and iterate Step 3 and Step 4 until convergence.

Table 1. Symmetric difference ($\boldsymbol{\Delta}$) between 10,000 trees sampled from the likelihood function via MCMC and the trees reconstructed by 5 methods. sub-EMGNJ means the implementation of subroutines, Step 3 and Step 4.

$\boldsymbol{\Delta}$	sub-EMGNJ	Saitou-Nei NJ	fastDNAml	DNAml(A)	DNAml(B)	TrExML
0	0	0	0	2	3608	0
2	77	0	0	1	471	0
4	3616	171	6	3619	5614	0
6	680	5687	5	463	294	5
8	5615	4134	3987	5636	13	71
10	12	8	5720	269	0	3634
12	0	0	272	10	0	652
14	0	0	10	0	0	5631
16	0	0	0	0	0	7

5. Computation

We implemented subroutines of the EMGNJ algorithm, Step 3 and Step 4 with $m = 4$ under the JC model. We plan that a full implementation of the algorithm will be released soon. We applied our implementation to find the phylogenetic tree for 21 *S-locus* receptor kinase (SRK) sequences from [14] involved in the self/nonself discriminating self-incompatibility system of the mustard family described in [12].

We sampled 10,000 trees from a Markov chain with stationary distribution proportional to the likelihood function by means of a Markov chain Monte Carlo (MCMC) algorithm implemented in `PHYBAYES` [1]. We then compared the tree topology of each tree generated by this MCMC method with that of the reconstructed trees via Step 3 and Step 4 in the EMGNJ method, Saitou-Nei NJ method, `fastDNAml`, `DNAml` from `PHYLIP` package by [7], and `TrExML` [19] under their respective default settings with the JC69 model. We used `treedist` from `PHYLIP` to compare two tree topologies. If the symmetric difference Δ between two topologies is 0, then the two topologies are identical. Larger Δ's are reflective of a larger distance between the two compared topologies. Table 1. summarizes the distance between a reconstructed tree and the MCMC samples from the normalized likelihood function. For example, the first two elements in the third row of Table 1. mean that 171 out of the 10,000 MCMC sampled trees are at a symmetric difference of 4 ($\Delta = 4$) from the tree reconstructed via Saitou-Nei NJ method (with pairwise distance). DNAml was used in two ways: DNAml(A) is a basic search with no global rearrangements, whereas DNAml(B) applies a broader search with global rearrangements and 100 jumbled inputs. The fruits of the broader search are reflected by the

accumulation of MCMC sampled trees over small Δ values from the DNAml(B) tree. Note that the tree topology for the reconstructed tree via Step 3 and Step 4 in the EMGNJ method and one via the GNJ method (`MJOIN` [11]) have the same tree topology.

6. Conclusion

We expect to improve the results using a full implementation of the EMGNJ method for two main reasons: (1) Even though currently we have an implementation with subroutines Step 3 and Step 4, the result is improved compared to Saitou-Nei NJ method and to `fastDNAml` as in Table 1. (2) Iterating Step 3 and Step 4 should improve the GTR rates.

We want to investigate the number of iterations until the output from the EMGNJ method converges. It is also of interest to study the results with different values of m.

7. Discussion

A potential problem is that the EM algorithm often get stuck in local maxima. We might be able to avoid this problem using Moore Rejection Sampling [16]. We sample data via the sampler for each subtree and we reconstruct trees with a set of samples under the JC69 model. Then we estimate the GTR rates and trees via the EMGNJ method. At the end, we compare the likelihood values of these trees and we take the tree with the biggest likelihood value. One notices that the process to reconstruct trees via the EMGNJ method from samples can be parallelizable.

Acknowledgement

AH acknowledges financial support from the Danish Research Council (Grant 21-04-0375).

References

[1] Aris-Brosou, S., How Bayes tests of molecular phylogenies compare with frequentist approaches. *Bioinformatics*, **19** (2003) 618–624.

[2] Buneman, P., The recovery of trees from measures of similarity, In Hodson, F.R., Kendall, D.G. and Tautu, P., editors, *Mathematics of the Archaeological and Historical Sciences*, Edinburgh University Press, Edinburgh, (1971) 387-395.

[3] Dempster, A.P., Laird, N.M. and Rubin, D.B., Maximum likelihood from incomplete data via the EM algorithm. *J.R. Statist. Soc. B* **39** (1977) 1-22.

[4] Faith, D.P. (1992). Conservation evaluation and phylogenetic diversity, *Biological Conservation*, **61** 1-10.

[5] Felsenstein, J., Inferring phylogenies from protein sequences by parsimony, distance and likelihood methods. *Methods in Enzymology*, **266** (1996) 418-427.

[6] Felsenstein, J., Evolutionary trees from DNA sequences. *J. Mol. Evol.* **17** (1981) 368-376.

[7] Felsenstein, J., PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author, (2004) Department of Genome Sciences, University of Washington, Seattle.

[8] Galtier, N., Gascuel, O. and Jean-Marie, A., Markov models in Molecular Evolution, In Nielsen, R., editor, *Statistical Methods in Molecular Evolution*, Springer, New York, (2005) 3-24.

[9] Guttorp, P., *Stochastic modeling of scientific data.* (1995) Chapman and Hall, Suffolk, Great Britain.

[10] Hobolth, A. and Jensen, J.L., Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Statistical applications in Genetics and Molecular Biology* (2005), **4**, 18.

[11] Levy, D., Yoshida, R. and Pachter, L., Neighbor Joining with Phylogenetic Diversity Estimates. Accepted for publication in *Molecular Biology and Evolution* (2005).

[12] Nasrallah, J.B., Recognition and rejection of self in plant reproduction. *Science*, **296** (2002) 305–308.

[13] Pachter, L. and Speyer, D., Reconstructing trees from subtree weights. *Applied Mathematics Letters*, **17** (2004) 615-621.

[14] Sainudiin, R., Wong, S.W., Yogeeswaran, K., Nasrallah, J., Yang, Z., and Nielsen, R., Detecting site-specific physicochemical selective pressures: applications to the class-I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *Journal of Molecular Evolution* in press (2005).

[15] Sainudiin, R. and Yoshida, R., Applications of Interval Methods to Phylogenetic trees. *Algebraic Statistics for Computational Biology* edited by Pachter, L. and Sturmfels, B., (2005) Cambridge University Press.

[16] Sainudiin, R., Machine Interval Experiments. PhD Thesis. (2005) Cornell University.

[17] Saitou, N. and Nei, M., The neighbor joining method: A new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution*, **4** (1987) 406-425.

[18] Studier, J.A. and Keppler, K.J., A note on the neighbor-joining method of Saitou and Nei, *Molecular Biology and Evolution*, **5** (1988) 729-731.

[19] Wolf, M.J., Easteal, S., Kahn, M., McKay, B.D., and Jermiin, L.S., TrExML: A maximum likelihood program for extensive tree-space exploration. *Bioinformatics*, **16** (2000) 383–394.

[20] Yap, V.B. and Speed, T.P., Modeling DNA base substitution in large genomic regions from two organisms. *J. Mol. Evol.* **58** (2004) 12-18.

Entry Form for the Proceedings

8. Title of the Paper

Maximum likelihood estimation of phylogenetic tree and substitution rates via generalized neighbor-joining and the EM algorithm

9. Author(s)

We are the authors of this paper.

Author No. 1
- Full Name: Asger Hobolth
- First Name: Asger
- Middle Name:
- Surname: Hobolth
- Initialized Name: A. Hobolth
- Affiliation: North Carolina State University, North Carolina, USA
- E-Mail: asger@statgen.ncsu.edu
- Ship the Proceedings to: BRC - Bioinformatics Research Center, 1500 Partners II, 840 Main Campus Drive North, Carolina State University, Raleigh North Carolina 27606, USA

Author No. 2
- Full Name: Ruriko Yoshida
- First Name: Ruriko
- Middle Name:
- Surname: Yoshida
- Initialized Name: R. Yoshida
- Affiliation: Duke University, North Carolina, USA
- E-Mail: ruriko@math.duke.edu
- Ship the Proceedings to: Mathematics Department, Duke University, Box 90320, Durham North Carolina 27708-0320, USA