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

Ruriko Yoshida
Dept. of Mathematics Duke University

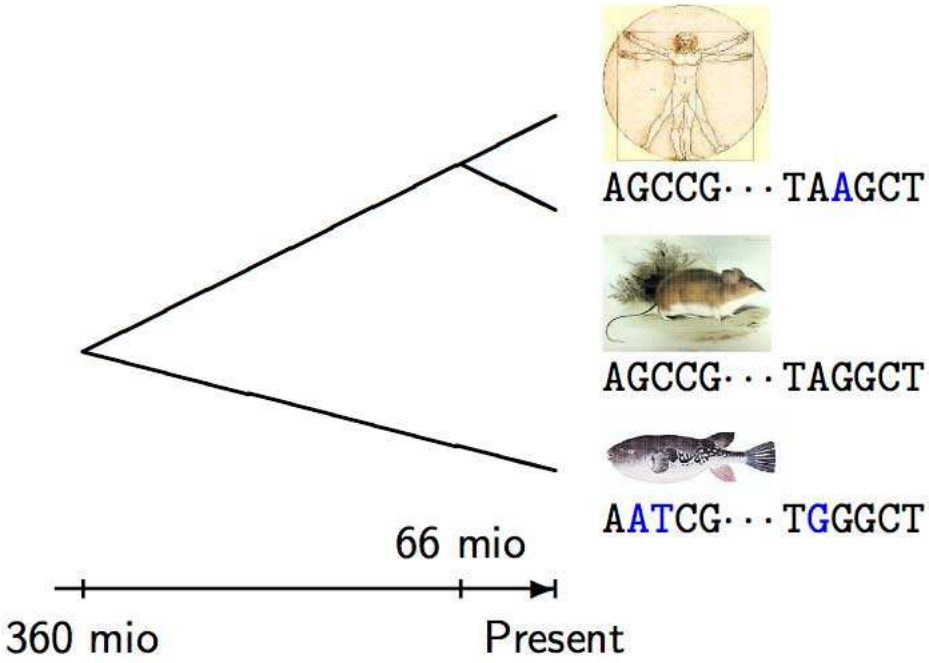
Joint work with Asger Hobolth

www.math.duke.edu/~ruriko

November 30th, 2005

Phylogeny

Phylogenetic trees describe the evolutionary relations among groups of organisms.



Why we care?

- Be able to analyze changes occurred in evolution of different species.
- Helps predict which species might have similar functions.
- Predicts changes occurring in rapid changing species, such as HIV virus.

Application

We would like to assemble the fungi tree of life.

Francois Lutzoni and Rytas Vilgalys Department of Biology, Duke University

$1500+$ fungal species

`http://ocid.nacse.org/research/aftol/about.php`

How can we reconstruct?

We can reconstruct a phylogenetic tree from DNA sequences via:

- **The maximum likelihood estimation (MLE) method**: it describes an evolution in terms of a discrete-state continuous-time Markov process. The substitution rate matrix can be estimated using the **expectation maximization (EM) algorithm**.

- **Distance based methods**: it computes pair-wise distances, which can be computed easily, and combinatorially reconstruct a tree. The most popular method is the **neighbor-joining (NJ) method**.

However

The MLE method: an exhaustive search for the ML phylogenetic tree is computationally prohibitive for large data sets.

The NJ method: sequence information is reduced. Especially, the NJ phylogenetic tree for large data sets loses so much sequence information.

Goal:

- Want an algorithm for simultaneous substitution rate estimation and phylogenetic tree reconstruction combining the MLE method and the NJ method.

- Want to apply for reconstructing a tree for large data sets.

The EMGNJ algorithm

The GNJ method: in 2005, Levy, Y., and Pachter introduced the **generalized neighbor-joining (GNJ) method**, which reconstructs a phylogenetic tree based on comparisons of subtrees rather than pairwise distances

- The GNJ method uses more sequence information: the resulting tree should be more accurate than the NJ method.

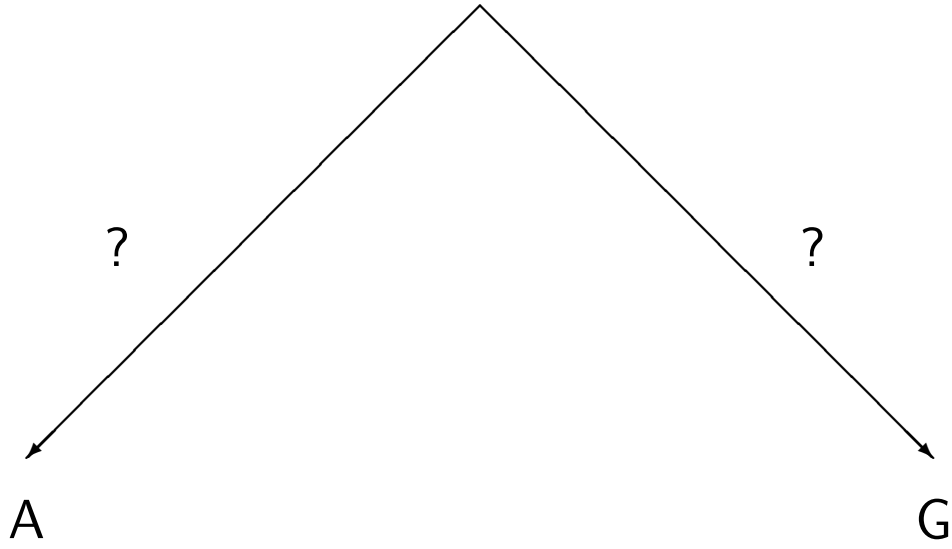
- The computational time: polynomial in terms of the number of DNA sequences.

The EMGNJ algorithm: iterates between the EM algorithm for estimating substitution rates and the generalized NJ method for phylogenetic tree reconstruction.

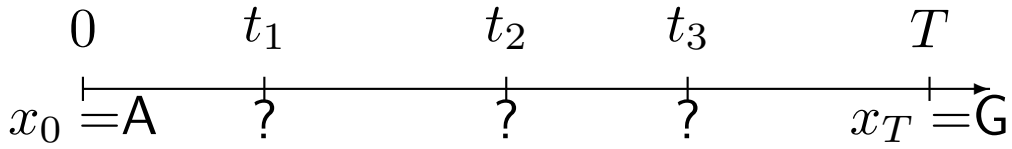
The EM algorithm

Pairwise sequences

Suppose we have a pair of sequences at a single site such that:

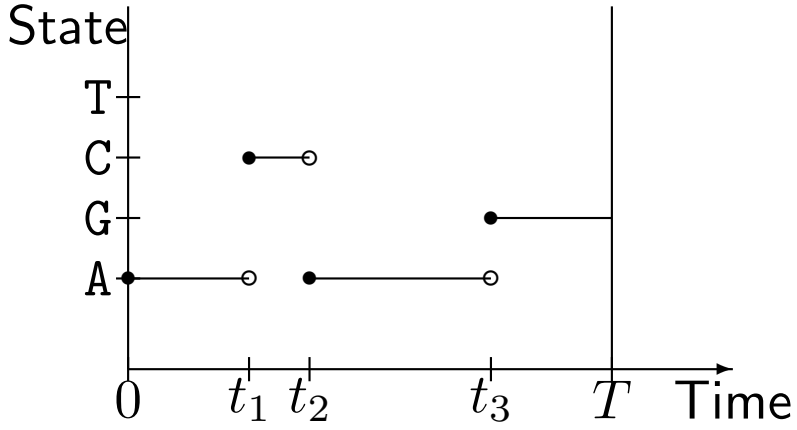


Assuming time reversibility....



Complete observation

Suppose we have the **complete observation** of continuous time Markov chain $x = \{x_t : 0 \leq t \leq T\}$ on the state space $\Sigma = \{A, C, G, T\}$.



We wish to estimate the substitution rate matrix Q using Maximum Likelihood.

Let $Q = (Q(a,b))_{a,b\in\Sigma} \in \mathbb{R}^{n\times n}$.

The waiting time in a state a is exponentially distributed with parameter $-Q(a,a)$ and the density for an exponential distribution with parameter r is $r\exp(-rt)$.

The prob. of substituting to state b different from a is: $-Q(a,b)/Q(a,a)$.

So the density for a substitution of a with b at time t is:

$$(-\mathbf{Q(a,a)})\exp(\mathbf{Q(a,a)t})\frac{\mathbf{Q(a,b)}}{(-\mathbf{Q(a,a)})} = \mathbf{Q(a,a)}\exp(\mathbf{Q(a,a)t})\frac{\mathbf{Q(a,b)}}{\mathbf{Q(a,a)}}$$
$$= \mathbf{Q(a,b)}\exp(\mathbf{Q(a,a)t}).$$

Thus, the likelihood for the complete observation is given by

$$L(Q,x) = Q(A,A)e^{Q(A,A)t_1}\frac{Q(A,C)}{Q(A,A)}Q(C,C)e^{Q(C,C)(t_2-t_1)}\frac{Q(C,A)}{Q(C,C)}$$

$$\times Q(A,A)e^{Q(A,A)(t_3-t_2)}\frac{Q(A,G)}{Q(A,A)}e^{Q(G,G)(T-t_2)}$$

$$= e^{Q(A,A)(t_1+(t_3-t_2))+Q(G,G)(T-t_2)+Q(C,C)(t_2-t_1)}Q(A,G)Q(A,C)Q(C,A)$$

$$= e^{Q(A,A)T(A)+Q(G,G)T(G)+Q(C,C)T(C)}$$

$$\times Q(A,G)^{N(A,G)}Q(A,C)^{N(A,C)}Q(C,A)^{N(C,A)}.$$

where $T(a)$ is the total time spent in state a and $N(a,b)$ is the number of substitutions of a with b.

More generally...

If Q is parametrized by θ, $Q = Q_\theta$, and with $x = \{x(t) : 0 \leq t \leq T\}$, the MLE problem for a complete observation is:

$$\max \quad \mathbf{L}(\theta; \mathbf{x}) \;=\; \left[\prod_{\mathbf{a}\in\mathbf{\Sigma}} \prod_{\mathbf{b}\neq\mathbf{a}} \mathbf{Q}_\theta(\mathbf{a},\mathbf{b})^{\mathbf{N}(\mathbf{a},\mathbf{b})}\right] \left[\prod_{\mathbf{a}\in\mathbf{\Sigma}} \exp(\mathbf{Q}_\theta(\mathbf{a},\mathbf{a}))^{\mathbf{T}(\mathbf{a})}\right]$$

such that $\theta \in \Theta$.

The log-likelihood for a complete observation becomes

$$\log \mathbf{L}(\theta; \mathbf{x}) = \sum_{\mathbf{a}\in\mathbf{\Sigma}} \mathbf{T}(\mathbf{a})\mathbf{Q}_\theta(\mathbf{a},\mathbf{a}) + \sum_{\mathbf{a}\in\mathbf{\Sigma}} \sum_{\mathbf{b}\neq\mathbf{a}} \mathbf{N}(\mathbf{a},\mathbf{b}) \log \mathbf{Q}_\theta(\mathbf{a},\mathbf{b}).$$

The GTR model

Consider the general time reversible (GTR) model.

Let $\pi_a,\ a \in \Sigma,\ \sum_a \pi_a = 1$, be the stationary distribution of the Markov chain.

The GTR model has substitution rate matrix:

$$Q_\theta = \begin{bmatrix} \cdot & \theta_{AG}\pi_G & \theta_{AC}\pi_C & \theta_{AT}\pi_T \\ \theta_{AG}\pi_A & \cdot & \theta_{GC}\pi_C & \theta_{GT}\pi_T \\ \theta_{AC}\pi_A & \theta_{GC}\pi_G & \cdot & \theta_{CT}\pi_T \\ \theta_{AT}\pi_A & \theta_{GT}\pi_G & \theta_{CT}\pi_C & \cdot \end{bmatrix}$$

where the diagonal elements are such that each row sums to zero.

The 6 unknown parameters are $\theta = (\theta_{AG}, \theta_{AC}, \theta_{AT}, \theta_{GC}, \theta_{GT}, \theta_{CT})$.

Missing data problem

Problem: if we only observe $x(0)$ and $x(T)$?

Note: the complete log-likelihood is maximized for

$$\theta^*_{\mathbf{ab}} = \frac{\mathbf{N(a,b)} + \mathbf{N(b,a)}}{\pi_{\mathbf{b}}\mathbf{T(a)} + \pi_{\mathbf{a}}\mathbf{T(b)}}, \quad \mathbf{a} \neq \mathbf{b}. \tag{1}$$

The EM algorithm:

1. (Expectation Step) Calculate $T(a)^* := E[T(a) : x(0), x(T)]$ and $N(a,b)^* := E[N(a,b) : x(0), x(T)]$.

2. (Maximization Step) Substitute $T(a)^*$ and $N(a,b)^*$ into Equation (1).

Iterate between Step 1 and Step 2 until convergence.

Thm. [Hobolth and Jensen]

Denote the transition matrix $P(t) = \exp(Qt)$.

- Time spent in state a

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

- Number of transitions between states a and b

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

Multiple sequences

We fix the tree topology for a tree T with n leaves. Note that there are $2n-1$ edges in T.

The single site complete log-likelihood becomes

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

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.

Apply the previous theorem and Felsenstein's peeling algorithm to solve the problem.

The GNJ method

MJOIN is available at `http://bio.math.berkeley.edu/mjoin/`.

Neighbor Joining Method

Def. We call a pair of two distinct leaves $\{i, j\}$ a **cherry** if there is exactly one intermediate node on the unique path between i and j.

Let $D(ij)$ be a pairwise distance between i and j.

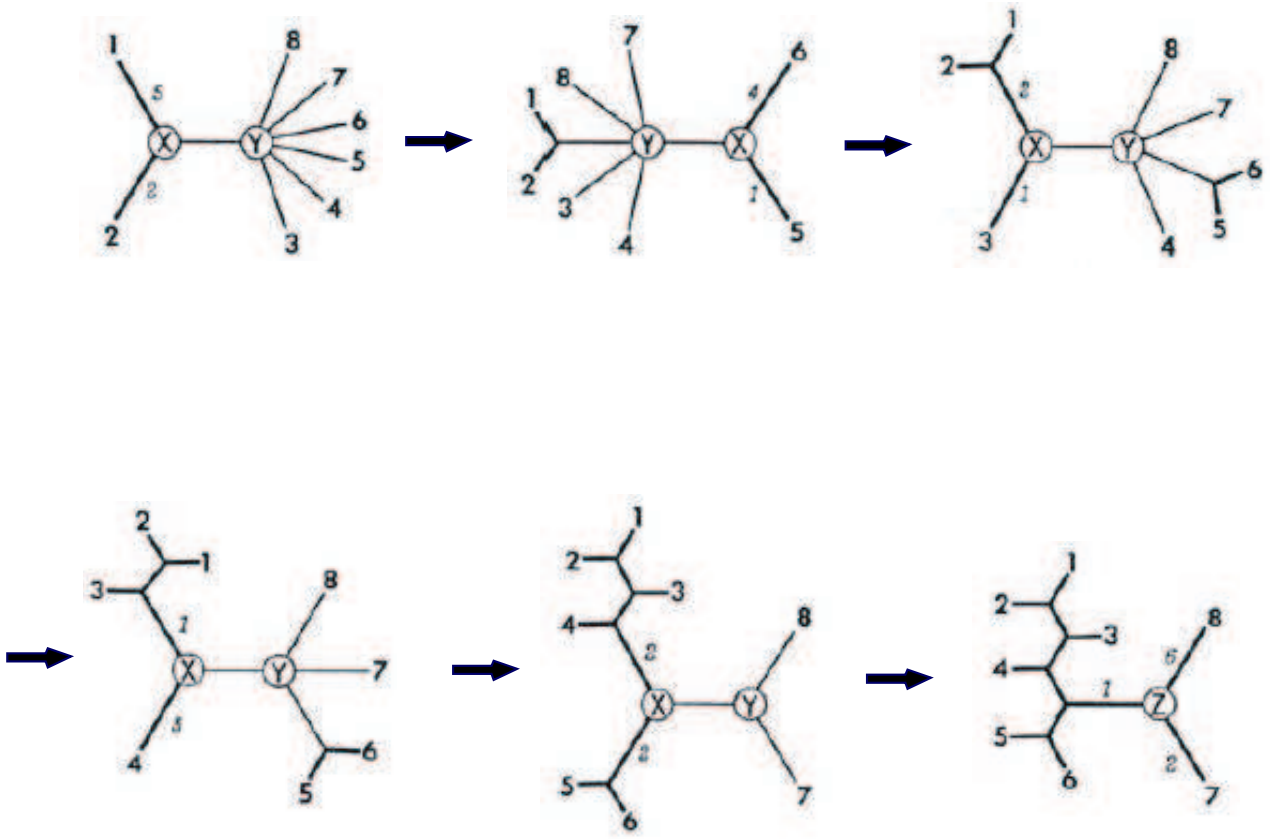
Thm. [Saitou-Nei and Studier-Keppler]

Let $A \in \mathbb{R}^{n \times n}$ such that $A_{ij} = D(ij) - (r_i + r_j)/(n-2)$, where $r_i := \sum_{k=1}^{n} D(ik)$. $\{i^*, j^*\}$ is a cherry in T if it minimizes A_{ij}.

Neighbor Joining Method:

Idea. Initialize a star-like tree. Then find a cherry $\{i, j\}$ and computing branch length from the interior node x to i and from x to j. Repeat this process recursively until we find all cherries.

Neighbor Joining Method



The GNJ method

- Extended the Neighbor Joining method with the total branch length of m-leaf subtrees.
- Increasing $2 \leq m \leq n-2$, a reconstructed tree from our method gets closer to the true tree.
- If $m = 3$, the time complexity of our new method is $O(n^3)$, which is the same as the Neighbor Joining with pairwise distances and a tree reconstructed by our method is more accurate than the one with pairwise distances.

Note: If $m = 2$, then our method is the Neighbor Joining method with pairwise distances.

Notation and Definitions

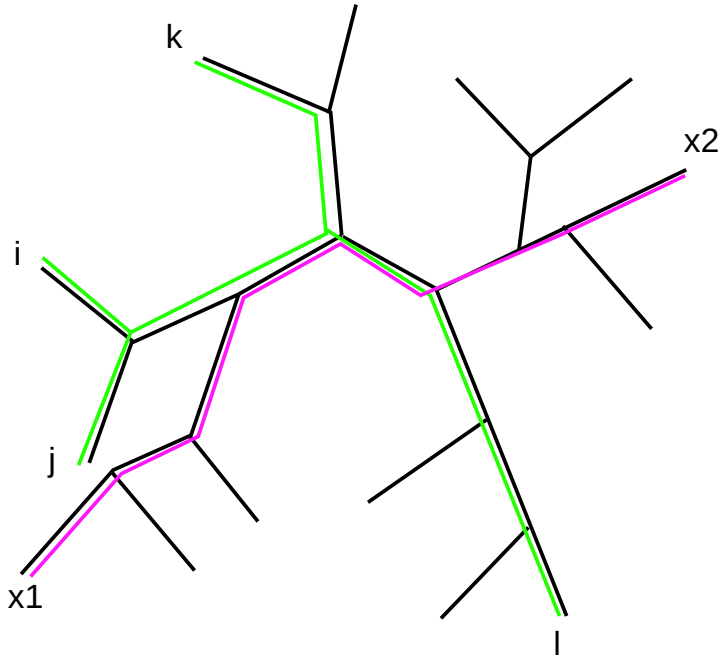
Notation. Let $[n]$ denote the set $\{1, 2, ..., n\}$ and $\binom{[n]}{m}$ denote the set of all m-element subsets of $[n]$.

Def. A **m-dissimilarity map** is a function $D : \binom{[n]}{m} \to \mathbb{R}_{\geq 0}$.

In the context of phylogenetic trees, the map $D(i_1, i_2, ..., i_m)$ measures the weight of a subtree that spans the leaves $i_1, i_2, ..., i_m$.

Denote $D(i_1 i_2 \dots i_m) := D(i_1, i_2, ..., i_m)$.

Weights of Subtrees in T



$D(ijkl)$ is the total branch length of the subtree in green. Also $D(x_1x_2)$ is the total branch length of the subtree in pink and it is also a pairwise distance between x_1 and x_2.

Thm. [Levy, Y., Pachter] Let D_m be an m-dissimilarity map on n leaves of a tree T, $D_m : \binom{[n]}{m} \to \mathbb{R}_{\geq 0}$ corresponding m-subtree weights, and define

$$\mathbf{S(ij)} := \sum_{\mathbf{X} \in \binom{[n] \setminus \{i,j\}}{m-2}} \mathbf{D_m(ijX)}.$$

Then $S(ij)$ is a tree metric.

Furthermore, if T' is based on this tree metric $S(ij)$ then there is an invertible linear map between their edge weights.

Note. This means that if we reconstruct T', then we can reconstruct T.

Computing edge weights on T

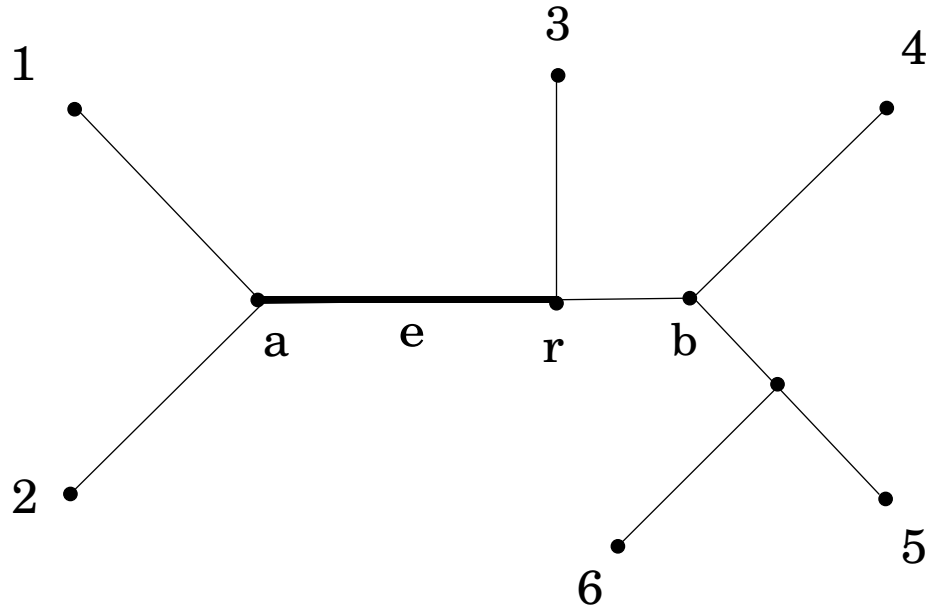
Lemma 1. [Levy, Y., Pachter] If e is an internal edge of T (equivalently T'), then

$$w_{T'}(e) = \frac{1}{2}[\binom{|L_1(e)|-2}{m-2} + \binom{|L_2(e)|-2}{m-2}]w_T(e)$$

where $L_1(e)$ and $L_2(e)$ are the two leaf sets of $T - e$.

For an edge e and a leaf i, $L_i(e)$ denotes the set of leaves in $T - e$ that are in the same connected component as i.

If $i = 3$, then $L_3(e) = \{3, 4, 5, 6\}$.



Lemma 2. [Levy, Y., Pachter] Denote the edges adjacent to the leaves by $e_1, \ldots, e_n$.

Let $C_i = \sum_{e \in int(E(T))} \left(\binom{n-2}{m-2} - \binom{|L_i(e)|-2}{m-2} \right) w_T(e)$. Then

$$\begin{pmatrix} w_T(e_1) \\ \vdots \\ w_T(e_n) \end{pmatrix} = A^{-1} \begin{pmatrix} 2w_{T'}(e_1) - C_1 \\ \vdots \\ 2w_{T'}(e_n) - C_n \end{pmatrix},$$

where $A^{-1} = \frac{1}{2\binom{n-2}{m-2}} \left(\mathbf{I} - \frac{m-2}{(m-1)(n-2)} \mathbf{J} \right)$.

Neighbor Joining with Subtree Weights

Input: n DNA sequences and an integer $2 \leq m \leq n-2$.

Output: A phylogenetic tree T with n leaves.

1. Compute all m-subtree weights via the maximum likelihood.

2. Compute $S(ij)$ for each pair of leaves i and j.

3. Apply Neighbor Joining method with a tree metric $S(ij)$ and obtain additive tree T'.

4. Using a one-to-one linear transformation, obtain a weight of each internal edge of T and a weight of each leaf edge of T.

Complexity

Lemma. [Levy, Pachter, Y.] If $m \geq 3$, the time complexity of this algorithm is $O(n^m)$, where n is the number of leaves of T and if $m = 2$, then the time complexity of this algorithm is $O(n^3)$.

Sketch of Proof: If $m \geq 3$, the computation of $S(ij)$ is $O(n^m)$ (both steps are trivially parallelizable). The subsequent neighbor-joining is $O(n^3)$ and edge weight reconstruction is $O(n^2)$. If $m = 2$, then the subsequent neighbor-joining is $O(n^3)$ which is greater than computing $S(ij)$. So, the time complexity is $O(n^3)$.

Note: The running time complexity of the algorithm is $O(n^3)$ for both $m = 2$ and $m = 3$.

Cherry Picking Theorem

Thm. [Levy, Pachter, Y.] Let T be a tree with n leaves and no nodes of degree 2 and let m be an integer satisfying $2 \leq m \leq n-2$. Let $D : \binom{[n]}{m} \to \mathbb{R}_{\geq 0}$ be the m-dissimilarity map corresponding to the weights of the subtrees of size m in T. Suppose we have:

$$\mathbf{B_D(ij)} = \left(\frac{\mathbf{n-2}}{\mathbf{m-1}}\right) \sum_{\mathbf{X} \in \binom{[n]\setminus\{i,j\}}{m-2}} \mathbf{D(ijX)} - \sum_{\mathbf{X} \in \binom{[n]\setminus\{i\}}{m-1}} \mathbf{D(iX)} - \sum_{\mathbf{X} \in \binom{[n]\setminus\{j\}}{m-1}} \mathbf{D(jX)}.$$

If $\{i^*, j^*\}$ is a pair such that $B_D(i^*j^*)$ is a minimal element of the matrix then $\{i^*, j^*\}$ is a cherry in the tree T.

Note. The theorem by Saitou-Nei and Studier-Keppler is a corollary from Cherry Picking Theorem.

The EMGNJ method

Algorithm

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 the GNJ method under the JC69 model.

3. Estimate GTR substitution rates and edge lengths from current tree via the EM algorithm.

4. Reconstruct tree using the GNJ method and current GTR rates.

5. If likelihood is not improved return current tree and GTR rates; otherwise go to 3.

Simulation Results

We implemented subroutines of the EMGNJ algorithm, Step 3 and Step 4 with $m = 4$ under the JC model.

Find the phylogenetic tree for 21 *S-locus* receptor kinase (SRK) sequences involved in the self/nonself discriminating self-incompatibility system of the mustard family.

Symmetric difference ($\mathbf{\Delta}$) between $10,000$ trees sampled from the likelihood function via MCMC and the trees reconstructed by 5 methods.

DNAml(A) is a basic search with no global rearrangements, whereas DNAml(B) applies a broader search with global rearrangements and 100 jumbled inputs.

A = sub-routine of the EMGNJ method, B = Saitou-Nei NJ method, C = fastDNAml, D = DNAml(A), F = DNAml(B), and G = TrExML.

Δ	A	B	C	D	F	G
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

Questions??

Thank you....