

Ruriko Yoshida

Neighbor Joining with Subtree Weights

Ruriko Yoshida

Dept. of Mathematics Duke University

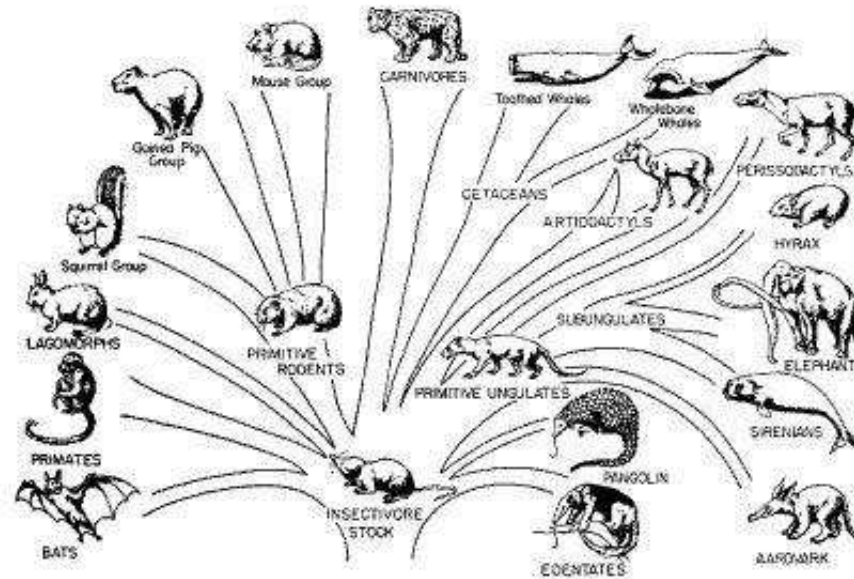
Joint work with Dan Levy and Lior Pachter

www.math.duke.edu/~ruriko

December 7th, 2004

Phylogeny

Phylogenetic trees illustrate the evolutionary relations among groups of organisms.



Why we care?

- We can analyze changes that have occurred in evolution of different species.
- Phylogenetic relations among different species help predict which species might have similar functions.
- We can predict changes occurring in rapid changing species, such as HIV virus.

Input Data for Phylogeny

- Numerical data
 - Distance between different species.
Each branch length represents evolutionary time along an edge of the tree.
We can derive distances from DNA sequences.
- Discrete characters
 - Each character has finite number of states
e.g. DNA = $\{A, T, C, G\}$.

Distance Based Methods

We reconstruct phylogenetic trees with distance based methods.

Input: Pair-wise distances from n many species.

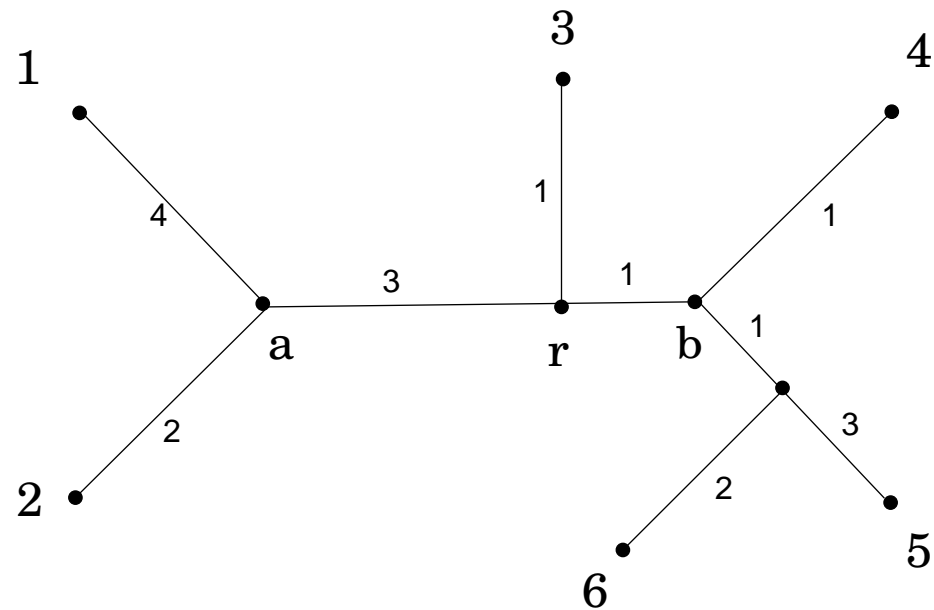
Output: An unrooted edge weighted binary tree with n leaves.

Distance based methods operate two steps:

1. Compute pair-wise distance between every pair of taxa.
2. With all pair-wise distances, compute tree topology and all branch length.

Distance Matrix

A distance matrix for a tree T is a matrix M whose entry M_{ij} stands for the mutation distance between i and j .



Distance Matrix

	1	2	3	4	5	6
1	0	6	8	9	12	11
2	6	0	6	7	10	9
3	8	6	0	3	6	5
4	9	7	3	0	5	4
5	12	10	6	5	0	5
6	11	9	5	4	5	0

Table 1: Distance matrix M for the example.

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

Definitions

Def. A distance matrix M is a **metric** iff M satisfies:

- Symmetric: $M_{ij} = M_{ji}$ and $M_{ii} = 0$.
- Triangle Inequality: $M_{ik} + M_{jk} \geq M_{ij}$.

Def. M is an **additive metric** iff there exists a tree T s.t.

- Every edge has a positive weight and every leaf is labeled by a distinct species in the given set.
- For every pair of i, j , M_{ij} = the sum of the edge weights along the path from i to j .

Also we call such T an **additive tree**.

If we want to reconstruct an additive tree T from an additive metric M , we can do in polynomial time on n .

Problem:

- A distance matrix M obtained from an alignment of DNA sequences is a non-additive metric.
- If M is not additive, finding the nearest additive metric \bar{M} is NP-hard (by Farach, Kannan, and Warnow).

We are interested in estimating the additive tree T of \bar{M} in polynomial time.

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 .

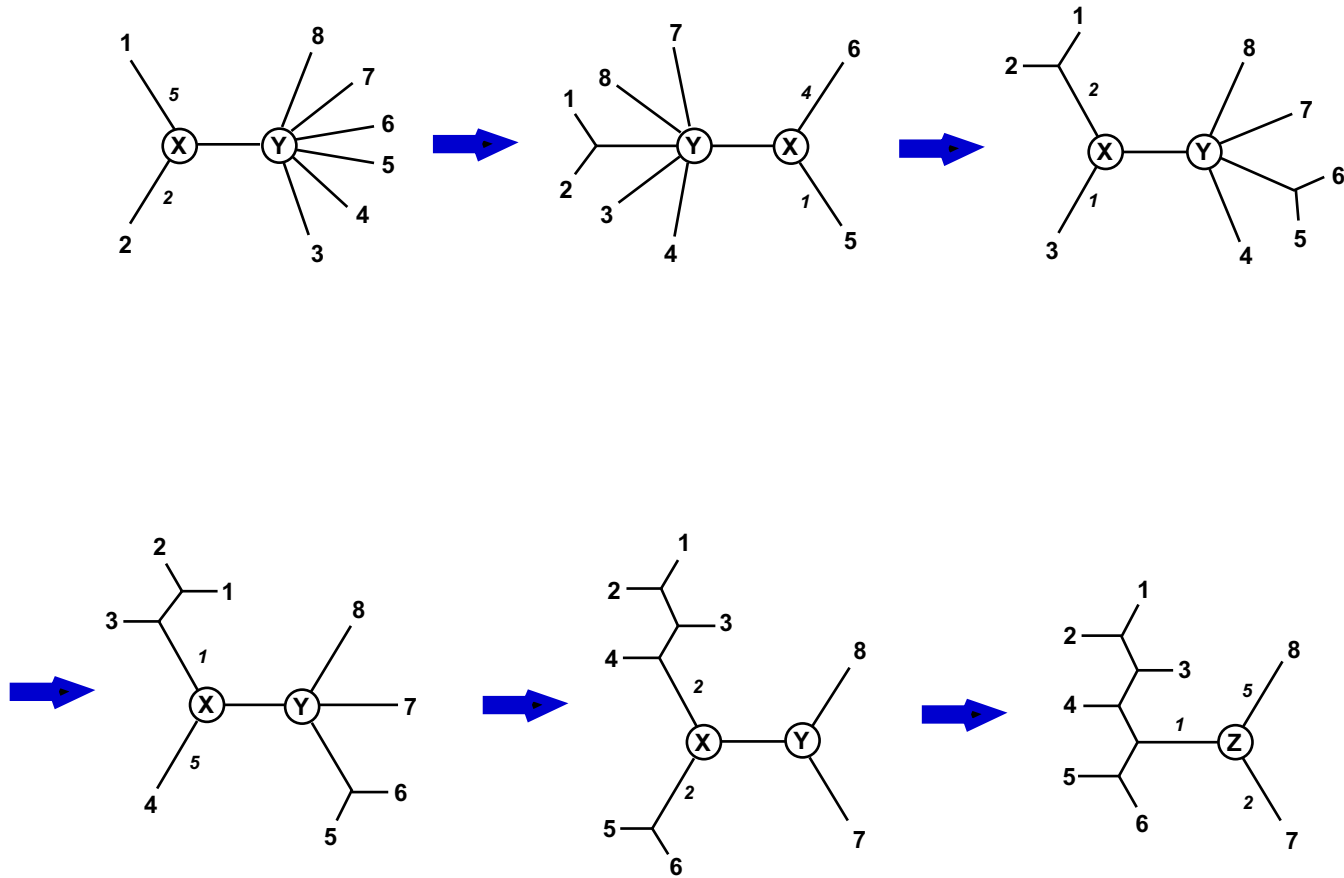
Thm. [Saitou-Nei and Studier-Keppler]

$\{i, j\}$ is a cherry if $A_{ij} = D(ij) - (r_i + r_j) / (n - 2)$, where $r_i := \sum_{k=1}^n D(ik)$, is minimal.

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



Advantages and Disadvantages

The most popular and widely used distance based method for reconstructing a phylogenetic tree.

Advantages:

- Fast (the time complexity of this algorithm is $O(n^3)$).
- Permits lineages with largely different branch lengths.

Disadvantages:

- Sequence information is reduced.
- Gives only one possible tree.

Neighbor Joining with Subtree Weights

- Extended the Neighbor Joining method with the weights of m -leaf subtrees.
- Increasing $2 \leq m \leq \frac{n+1}{2}$, reconstructed tree from our method gets closer to the additive tree of the nearest additive matrix.
- 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.

Notations and Definitions

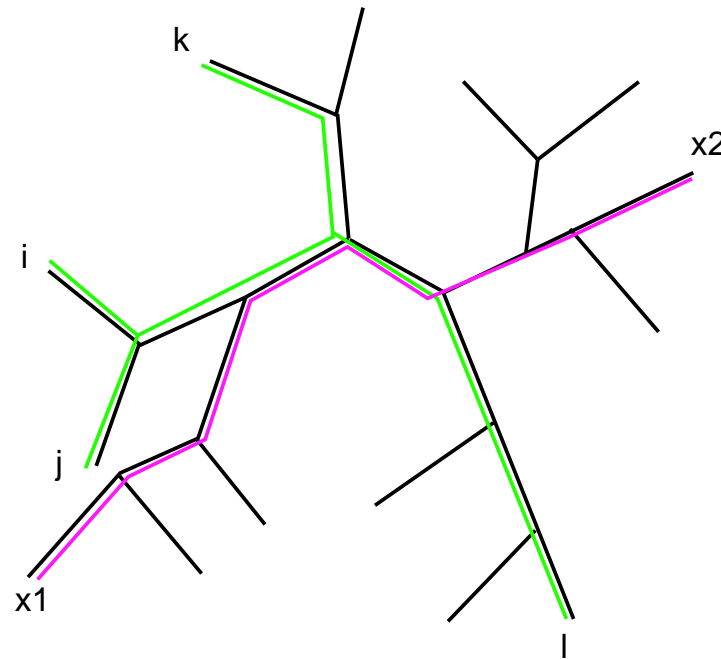
Notation. Let $[n]$ denote the set $\{1, 2, \dots, n\}$ and $\binom{[n]}{m}$ denote the set of all m -element subsets of $[n]$.

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

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

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

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 .

Main Theorem

Thm. [Levy, Pachter, Y.] Let D_m be an m -dissimilarity map on n leaves, $D_m : \binom{[n]}{m} \rightarrow \mathbb{R}_{\geq 0}$, and define

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

If the weights D_m correspond to m -subtree weights of a tree T then $S(ij)$ is a tree metric.

Furthermore, if T' is the additive tree corresponding to this tree metric then T' and T have the same tree topology and 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

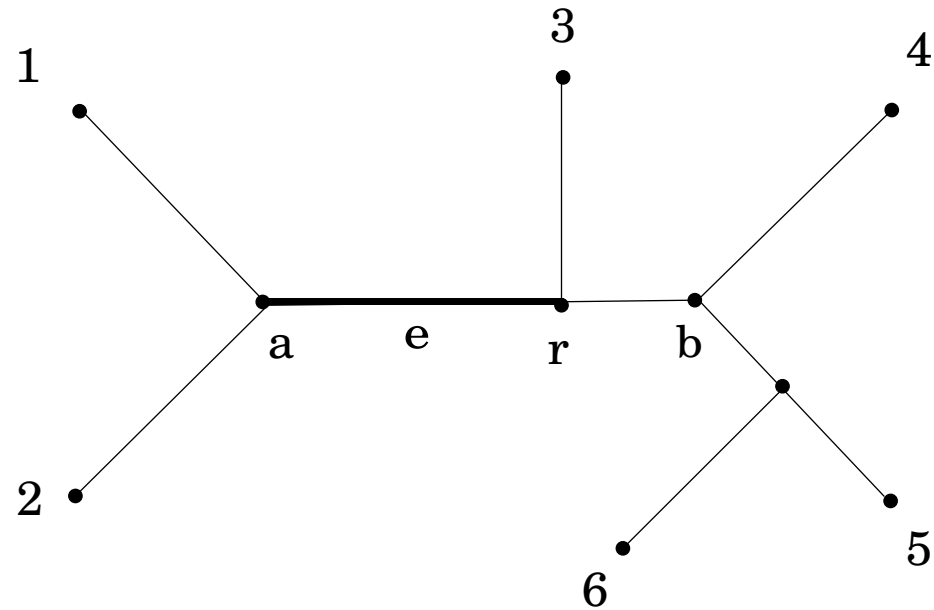
Lemma 1. [Levy, Pachter, Y.] If e is an internal edge of T (equivalently T'), then

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

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

For an edge $e \in E(T)$ 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, Pachter, Y.] Denote the edges adjacent to the leaves by e_1, \dots, e_n .

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

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 Lemma 1, obtain a weight of each internal edge of T .
5. Using Lemma 2, obtain 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 $m = 3$.

We need the following theorem and the main theorem follows.

Thm. (Cherry Picking Theorem) [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 $D : \binom{[n]}{m} \rightarrow \mathbb{R}_{\geq 0}$ be the m -dissimilarity map corresponding to the weights of the subtrees of size m in T . If $Q_D(ij)$ is a minimal element of the matrix

$$Q_D(ij) = \binom{n-2}{m-1} \sum_{X \in \binom{[n] \setminus \{i,j\}}{m-2}} D(ijX) - \sum_{X \in \binom{[n] \setminus \{i\}}{m-1}} D(iX) - \sum_{X \in \binom{[n] \setminus \{j\}}{m-1}} D(jX)$$

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.

Sketch of Proof for the Main Theorem

It is very easy to check that $M = (S(ij))$ is a tree metric by checking its symmetry and triangle inequality.

Then we set

$$A_{ij} = S(ij) - \frac{(r_i + r_j)}{(n - 2)},$$

where $r_i = \sum_{k=1}^n S(ik)$.

Notice that $r_i = \sum_{k=1}^n S(ik) = (m - 1) \sum_{Y \in \binom{[n] \setminus \{i\}}{m-1}} D(iY)$.

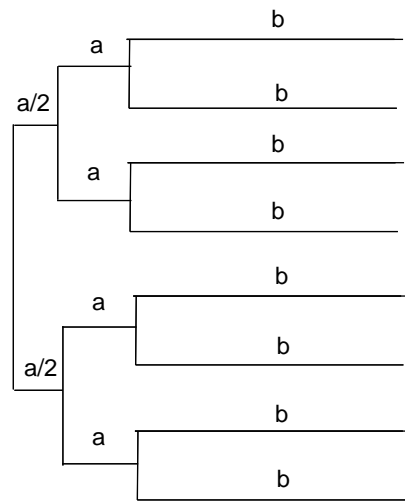
Therefore, $A_{ij} = Q_D(ij)$.

By the Cherry Picking Theorem, T and T' have the same tree topology.

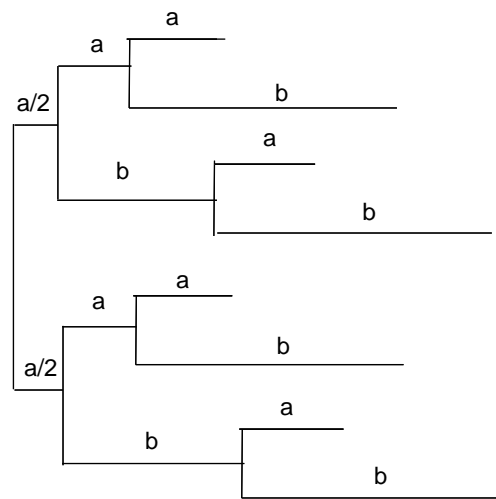
Simulation Results

Consider two tree models...

Modeled from Strimmer and von Haeseler.



T1



T2

We generate 500 replications with the Jukes-Cantor model via a software `evolver` from PAML package.

The number represents a percentage which we got the same tree topology.

l	a/b	m=2	m=3	m=4	fastDNAml
500	0.01/0.07	68.2	76.8	80.4	74.8
	0.02/0.19	54.2	61.2	73.6	55.6
	0.03/0.42	10.4	12.6	23.8	12.6
1000	0.01/0.07	94.2	96	97.4	96.6
	0.02/0.19	87.6	88.6	96.2	88
	0.03/0.42	33.4	35	52.4	33.6

Table 2: Success Rates for the model T_1 .

l	a/b	m=2	m=3	m=4	fastDNAml
500	0.01/0.07	84.4	86	85.6	88.4
	0.02/0.19	68.2	72	73.2	88.4
	0.03/0.42	18.2	29.2	36.2	87.4
1000	0.01/0.07	95.6	97.8	97.4	99.4
	0.02/0.19	88.4	89.6	93.4	99.8
	0.03/0.42	40	48.2	57.6	96.6

Table 3: Success Rates for the model T_2 .

Questions??

Paper and Software

Software package MJOIN is available.

Download at <http://bio.math.berkeley.edu/mjoin/>

You can download our paper at

<http://www.math.duke.edu/~ruriko/recomb.pdf>

Ruriko Yoshida

Thank you...