Partitioning the Sample Space on Five Taxa for the Neighbor Joining Algorithm

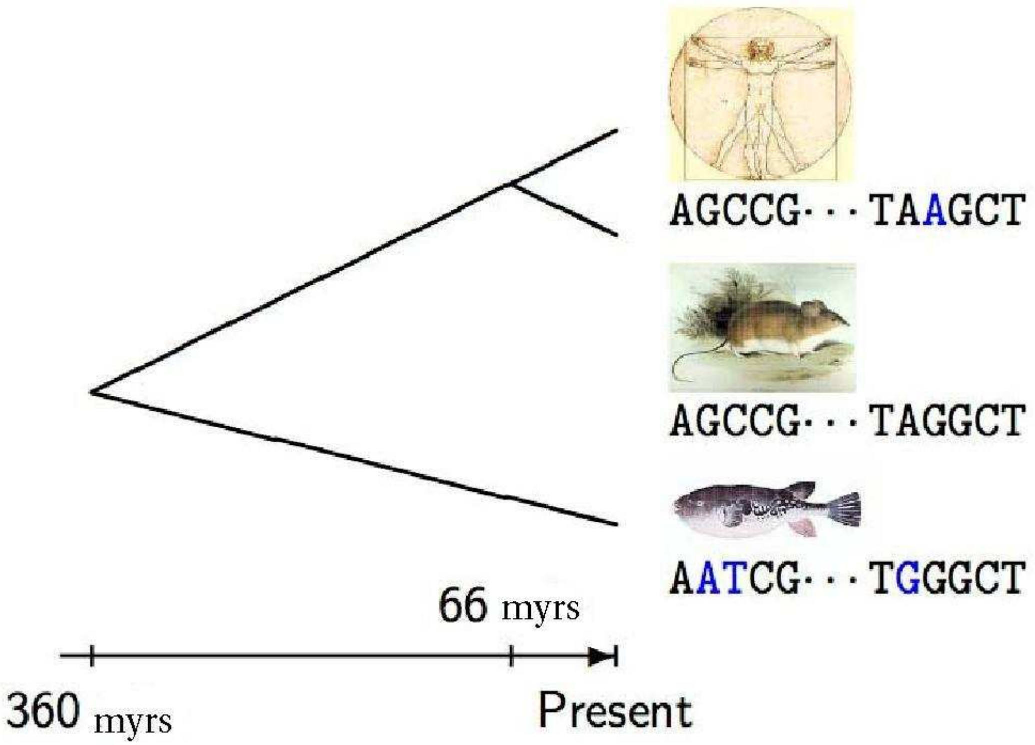
Ruriko Yoshida
Dept. of Statistics University of Kentucky

Joint work with Kord Eickmeyer

www.ms.uky.edu/~ruriko

Phylogeny

Phylogenetic trees describe 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.

Constructing trees from sequence data

“Ten years ago most biologists would have agreed that all organisms evolved from a single ancestral cell that lived 3.5 billion or more years ago. More recent results, however, indicate that this family tree of life is far more complicated than was believed and may not have had a single root at all.” (W. Ford Doolittle, (June 2000) *Scientific American*).

Since the proliferation of Darwinian evolutionary biology, many scientists have sought a coherent explanation from the evolution of life and have tried to reconstruct phylogenetic trees.

Methods to reconstruct a phylogenetic tree from DNA sequences include:

- **The maximum likelihood estimation (MLE) methods**: They describe evolution in terms of a discrete-state continuous-time Markov process. The substitution rate matrix can be estimated using the **expectation maximization (EM) algorithm**. (for eg. Dempster, Laird, and Rubin (1977), Felsenstein (1981)).

- **The Minimum Evolution (ME) method**: This is a **distance based method** and weighted Least Square method (the principle of Least Squares is a general method for estimating unknown parameters values so that error is minimized). It finds a closest additive metric from the given non-additive distance matrix with the smallest branch lengths (more biologically makes sense).

However

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

The ME method: This is an NP hard algorithm in terms of the number of taxa (Farach, Kannan, Warnow (1996), Rzhetsky and Nei (1993)).

Estimation: To solve the time complexity, we estimate the closest additive tree with smallest branch lengths.

Neighbor-joining (NJ) method: This is the most popular distance based method which computes a tree from all pair-wise distances obtained easily. It combinatorially estimates the ME tree (it is a greedy algorithm to find the ME tree) (Saito and Nei (1987), Studier and Keppler (1988)).

However again....

The NJ phylogenetic tree for large data sets loses so much sequence information and we do not know how well it performs with pairwise distances that are not tree metrics, especially when all pairwise distances are estimated via the MLE.

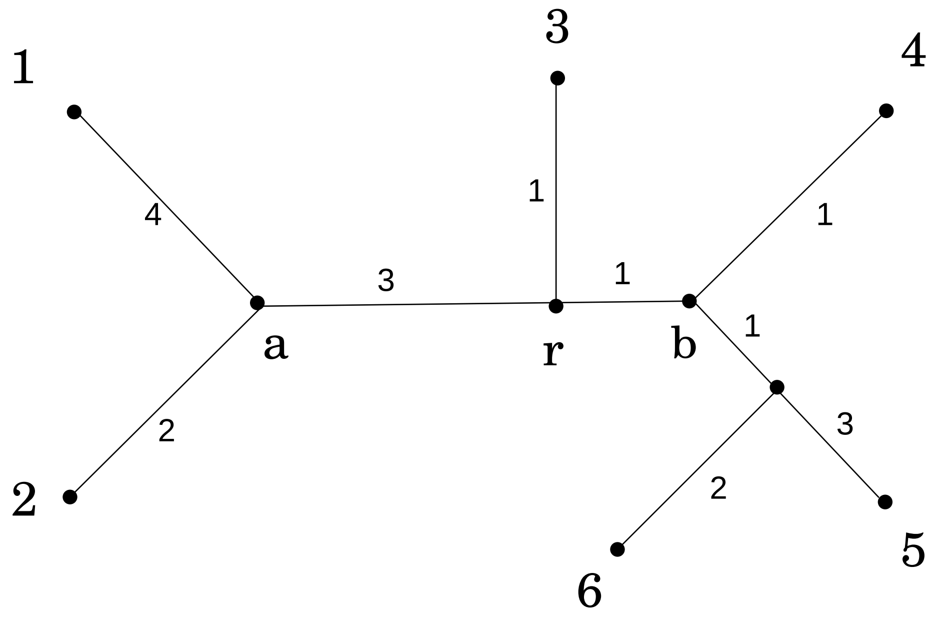
Goal:

- Analyze the behavior of the Neighbor Joining algorithm on five taxa.
- Using polyhedral geometry, partition the sample (data) space for estimation of a tree topology with five taxa into subspaces, within each of which the Neighbor Joining algorithm returns the same tree topology.

The NJ method

Distance Matrix

A **distance matrix** for a tree T is a matrix D whose entry D_{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 D for the example.

Definitions

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

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

Def. D 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, $D_{ij} =$ the sum of the edge weights along the path from i to j.

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

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.

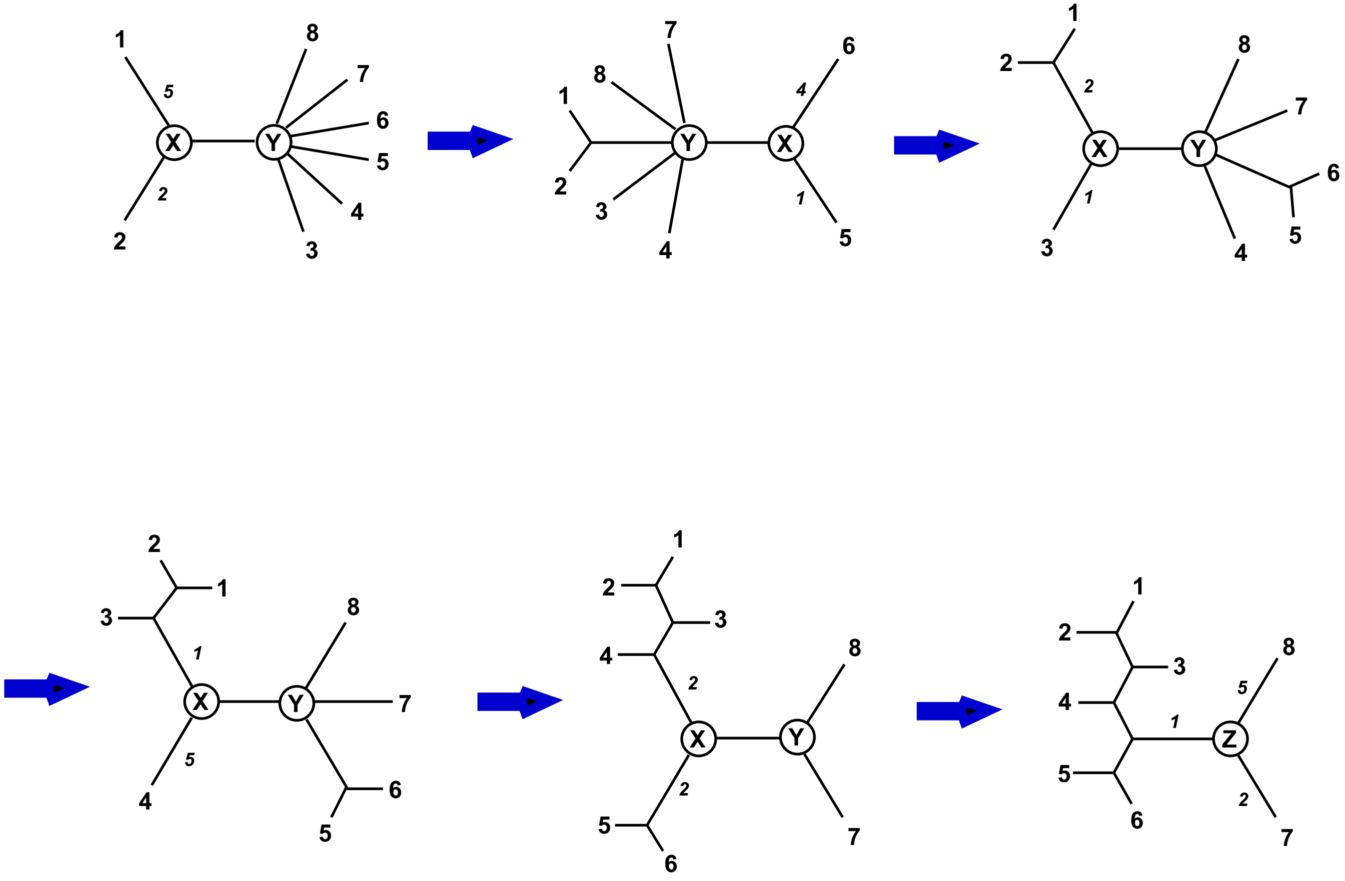
Thm. (Q-criterion) [Saitou-Nei, 1987 and Studier-Keppler, 1988]

Let $Q \in \mathbb{R}^{n \times n}$ such that $Q_{ij} = D_{ij} - (r_i + r_j)/(n-2)$, where $r_i := \sum_{k=1}^{n} D_{ik}$. $\{i^*, j^*\}$ is a cherry in T if $Q_{i^* j^*}$ is a minimum for all i and j.

Neighbor Joining Method:

Input. A tree matric D. **Output.** An additive tree T.
Idea. Initialize a star-like tree. Then find a cherry $\{i, j\}$ and compute 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 Q-criterion

The resulting matrix is again symmetric, and we can see it as a vector of dimension $m = \binom{n}{2}$ just like the input data. Moreover, the Q-criterion is obtained from the input data by a linear transformation:

$$\mathbf{q} = \mathbf{A^{(n)}d},$$

where $\mathbf{d}$ is a vector representation of D, $\mathbf{q}$ is a vector representation of Q,and the entries of the matrix $A^{(n)}$ are given by

$$\mathbf{A^{(n)}_{ij}} = \mathbf{A^{(n)}_{ab,cd}} = \begin{cases} n-4 & \text{if } i = j, \\ -1 & \text{if } i \neq j \text{ and } \{a,b\} \cap \{c,d\} \neq \emptyset, \\ 0 & \text{else,} \end{cases}$$

where $a > b$ is the row/column-index equivalent to i and likewise for $c > d$ and j.

Example

For $n = 4$ we have

$$A^{(4)} = \begin{pmatrix} 0 & -1 & -1 & -1 & -1 & 0 \\ -1 & 0 & -1 & -1 & 0 & -1 \\ -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & 0 & -1 & -1 & 0 & -1 \\ 0 & -1 & -1 & -1 & -1 & 0 \end{pmatrix}.$$

The Q-criterion:

find smallest $\mathbf{q_i}$ for $\mathbf{i = 1, \cdots, m}$ such that $\mathbf{q = A^{(4)}d}$.

Reducing the number of taxa

Suppose out of our n taxa $\{1, \ldots, n\}$, the first cherry to be picked is the $\binom{n}{2}$th cherry $\{n-1, n\}$, which we view as the new node number $n-1$.

The reduced pairwise distance matrix is one row and one column shorter than the original one. Explicitly,

$$\mathbf{d'_i} = \begin{cases} d_i & \text{for } 1 \le i \le \binom{n-2}{2} \\ \frac{1}{2}(d_i + d_{i+(n-2)} - d_{m-1}) & \text{for } \binom{n-2}{2} + 1 \le i \le \binom{n-1}{2} \end{cases}$$

We see that the reduced distance matrix depends linearly on the original one:

$$\mathbf{d}' = \mathbf{R}\mathbf{d},$$

with $R = (r_{ij}) \in \mathbb{R}^{(m-n+1)\times m}$, where

$$r_{ij} = \begin{cases} 1 & \text{for } 1 \leq i = j \leq \binom{n-2}{2} \\ 1/2 & \text{for } \binom{n-2}{2} + 1 \leq i \leq \binom{n-1}{2}, j = i \\ 1/2 & \text{for } \binom{n-2}{2} + 1 \leq i \leq \binom{n-1}{2}, j = i + n - 1 \\ -1/2 & \text{for } \binom{n-2}{2} + 1 \leq i \leq \binom{n-1}{2}, j = m \\ 0 & \text{else} \end{cases}$$

The process of picking cherries is repeated until there are only three taxa left, which are then joined to a single new node.

Shiftting Lemma

Note: There is an n-dimensional linear subspace of $\mathbb{R}^m$ which does not affect the outcome of NJ (Mihaescu et al, 2006). For a node a we define its *shift vector* $\mathbf{s}_a$ by

$$(\mathbf{s_a})_{\mathbf{b,c}} := \begin{cases} 1 & \text{if } a \in \{b, c\} \\ 0 & \text{else} \end{cases}$$

which represents a tree where the leaf a has distance 1 from all other leaves and all other distances are zero. The Q-criterion of any such vector is -2 for all pairs, so adding any linear combination of shift vectors to an input vector does not change the relative values of the Q-criteria.

The first step in cherry picking

After computing the Q-criterion $\mathbf{q}$, the NJ algorithm proceeds by finding the minimum entry of it, or, equivalently, the maximum entry of $-\mathbf{q}$. The set $cq_i \subset \mathbb{R}^m$ of all q-vectors for which q_i is minimal is given by

$$\begin{aligned} \mathbf{q} \in cq_i \quad & \Leftrightarrow \quad i = \arg\max(-e_j, A\mathbf{d}) \\ & \Leftrightarrow \quad i = \arg\max(-A^T e_j, \mathbf{d}) \\ & \Leftrightarrow \quad i = \arg\max(-Ae_j, \mathbf{d}) \quad \text{because } A \text{ is symmetric.} \end{aligned} \tag{1}$$

Therefore the set cd_i of all *parameter* vectors $\mathbf{d}$ for which the NJ algorithm will select cherry i in the first step is the normal cone at $-Ae_i$ to the polytope

$$\mathbf{P_n} := \mathrm{conv}\{-\mathbf{Ae_1}, \ldots, -\mathbf{Ae_m}\}. \tag{2}$$

The shifting lemma implies that the affine dimension of the polytope P_n is at most $m - n$.

Example for $n = 4$

In the case of four taxa, this reduces to a triangle with vertices

$$\mathbf{p}_0 = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{p}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{p}_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}.$$

The normal cones are bounded by three hyperplanes whose normal vectors are

$$\mathbf{n}_{01} = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}, \quad \mathbf{n}_{12} = \begin{pmatrix} 0 \\ -1 \\ 1 \\ 1 \\ -1 \\ 0 \end{pmatrix}, \quad \mathbf{n}_{20} = \begin{pmatrix} 1 \\ 0 \\ -1 \\ -1 \\ 0 \\ 1 \end{pmatrix}.$$

For example, the parameter region for which the pair of taxa 1 and 2 is chosen is defined by

$$cd_0 = \{\mathbf{x} \in \mathbb{R}^m \mid (\mathbf{n}_{01}, \mathbf{x}) > 0 \text{ and } (\mathbf{n}_{20}, \mathbf{x}) < 0\}.$$

The cone $C_{45,3}$

Since we can apply a permutation $\sigma \in S_5$ on taxa, without loss of generality, we suppose that the first cherry to be picked is the cherry with leaves 4 and 5. This is true for all input vectors $\mathbf{d}$ which satisfy

$$\mathbf{(h_{10,i}, d) \geq 0 \text{ for } i = 1, \ldots, 9,}$$

where the vector

$$\mathbf{h_{ij}^{(n)} := -A^{(n)}(e_i - e_j).}$$

Then, the set of all input vectors $\mathbf{d}$ for which the first picked cherry is 4-5 and the second one is 1-2:

$$C_{45,3} :=$$
$$\{\mathbf{d} \mid (\mathbf{h}_{10,i}, \mathbf{d}) \geq 0 \text{ for } i = 1, \ldots, 9, \text{ and } (\mathbf{r}_1 - \mathbf{r}_2, \mathbf{d}) \geq 0, (\mathbf{r}_1 - \mathbf{r}_3, \mathbf{d}) \geq 0\}$$

where $\mathbf{r}_1$, $\mathbf{r}_2$ and $\mathbf{r}_3$ are the first three rows of $-A^{(4)}R^{(5)}$.

The NJ cones

For $n = 5$, there is only one unlebeled tree and there are 15 lebeled trees. There are 30 cones in the 5 dimension (i.e. there are two cones per a lebeled tree).

- They do not form a fan.
- The union of cones $C_{12,3}$ and $C_{45,3}$ does not form a convex body (i.e. the union of two cones for one tree topology does not form a convex cone).

The edge radius

Theorem [Atteson, 1999]

Neighbor-joining has l_∞ radius $\frac{1}{2}$.

This means that if the distance estimates are at most half the minimal edge length of the tree away from their true value then the NJ algorithm will reconstruct the correct tree.

Theorem [Eickmeyer and Y, 2007]

Neighbor-joining has l_2 radius $\frac{1}{\sqrt{3}}$.

For $n = 6$

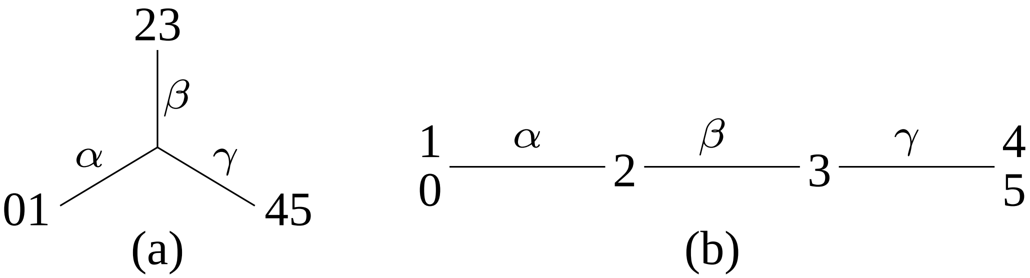


Figure 1: The two possible topologies for trees with six leaves, with edges connecting to leaves shrunk to zero.

There are three different classes of cones which cannot be mapped onto each other by the group action, C_I, C_{II}, C_{III}.

- **Type I**: $a, b, c, d, e, f \to a, b, c, d, (ef) \to a, b, (cd), (ef), \to$ Fig. 1(a)

- **Type II**: $a, b, c, d, e, f \to a, b, c, d, (ef) \to a, b, (cd), (ef)$ $\to cd - a - b - ef$ (like Fig 1(b), but different labels)

- **Type III**: $a, b, c, d, e, f, \to a, b, c, d, (ef) \to a, b, c, (d(ef))$ $\to ab - c - d - ef$ (exactly as in Fig 1(b))

	C_{I}	C_{II}	C_{III}
stabilizer	$\langle(12), (34), (56)\rangle$	$\langle(12), (56)\rangle$	$\langle(12), (56)\rangle$
size of stabilizer	8	4	4
number of cones	90	180	180
cones giving same labeled topology	6	2	2

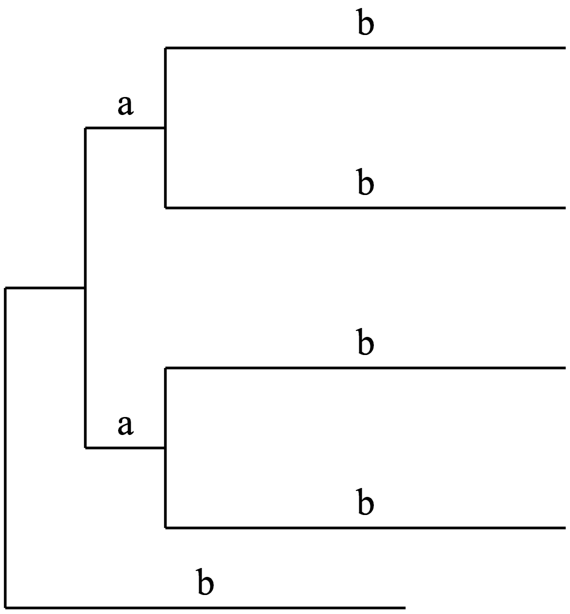
Simulation Results

With the Juke Cantor and Kimura 2 parameter models.

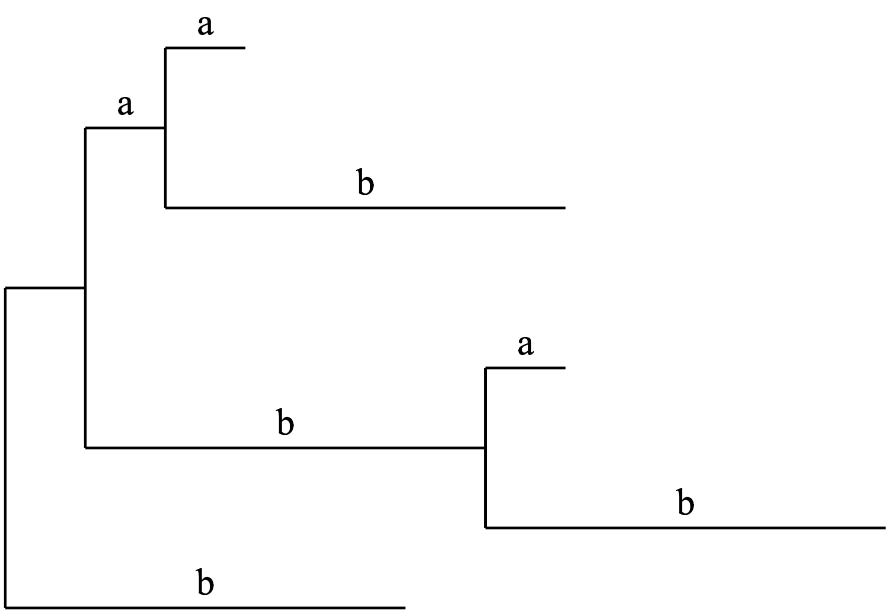
Consider two tree models...

Modeled from Strimmer and von Haeseler.

$\mathbf{T_1}$



$\mathbf{T_2}$



We generate $10,000$ replications at the edge length ratio, a/b $=$ 0.03/0.42 for sequences of length 500BP with the Jukes-Cantor and Kimura 2 parameter models via a software `evolver` from `PAML` package.

For each set of 5 sequences, we compute first pairwise distances via the heuristic MLE method using a software `fastDNAml`. To compute cones, we used `MAPLE` and `polymake`.

We say an input vector (distance matrix) is **correctly classified** if the vector locates in one of the cones where the vector representation of the tree metric (noiseless input) lies. We say an input vector is **incorrectly classified** if the vector locates in the complement of the cones where the vector representation of the tree metric lies.

For distance matrices which are correctly classified by the NJ algorithm, we compute the minimum distance to any cone giving a different tree topology.

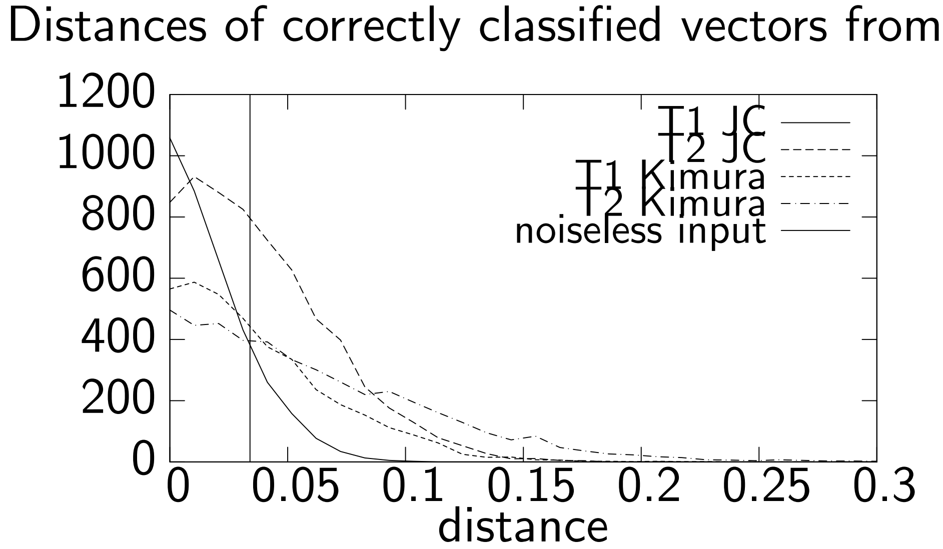


Figure 2: Distances of correctly classified input vectors from the closest correctly classified vector.

Mean and variance of the distances of correctly classified vectors from the nearest misclassified vector.

	JC		**Kimura2**	
	T1	T2	T1	T2
# of cases	3,581	6,441	3,795	4,467
Mean	0.0221	0.0421	0.0415	0.0629
Variance	$2.996 \cdot 10^{-4}$	$9.032 \cdot 10^{-4}$	$1.034 \cdot 10^{-3}$	$2.471 \cdot 10^{-3}$

For input vectors to which the NJ algorithm answers with a tree topology different from the correct tree topology, we compute the distances to the two cones for which the correct answer is given and take the minimum of the two. The bigger this distance is, the further we are off.

Distances of misclassified input vectors from closest correctly classified vector

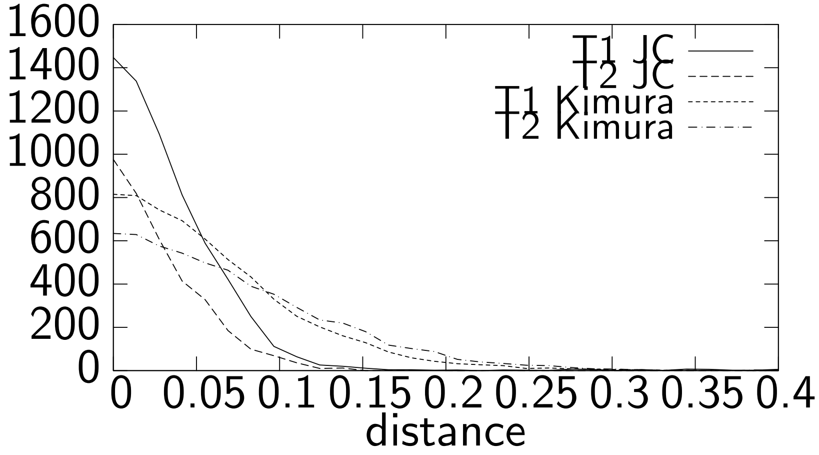


Figure 3: Distances of correctly incorrectly classified input vectors from the closest incorrectly classified vector.

Mean and variance of the distances of misclassified vectors to the nearest correctly classified vector.

	JC		**Kimura2**	
	T1	T2	T1	T2
# of cases	6,419	3,559	6,205	5,533
Mean	0.0594	0.0331	0.0951	0.0761
Variance	0.0203	$7.39 \cdot 10^{-4}$	0.0411	$3.481 \cdot 10^{-3}$

Future work

- Study the intersections of cones for $n = 5$ and $n = 6$ more closely.
- We have computed the cones for general n.
- Compare the NJ cones with the cones for the ME method (with Lior Pachter and Peter Huggins).

Thank you....

The preprint is available at math.CO/0703081.