Neighbor-Joining Algorithm:

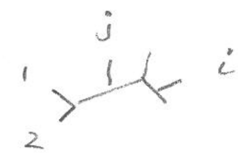
Building on the 4-point condition. Before beginning, keep in mind

- want NJ to return correct tree on pairwise dissimilarities from a tree metric (1)
- want reasonable way to average distances when no underlying tree metric

Notation: $|X| = N$ N taxa S_i, $i = 1, 2, \ldots, N$

$d_{ij} = d(S_i, S_j) =$ pairwise dissimilarity between S_i, S_j

With (1) in mind, assume 1,2 in cherry on tree, then for all $i, j = 3, 4, \ldots, N$



$$d_{12} + d_{ij} < d_{1i} + d_{2j}$$

Now fix i, and sum over all j

$$\sum_{\substack{j=3 \\ j \neq i}}^{N} (\underbrace{d_{12}}_{\text{fixed}} + d_{ij}) < \sum_{\substack{j=3 \\ j \neq i}}^{N} (\underbrace{d_{1i}}_{\text{fixed}} + d_{2j})$$

$$\Rightarrow (N-3)\, d_{12} + \sum_{\substack{j=3 \\ j \neq i}}^{N} d_{ij} < (N-3)\, d_{1i} + \sum_{\substack{j=3 \\ j \neq i}}^{N} d_{2j}$$

Add $d_{12} + d_{i1} + d_{i2}$ to both sides

$$(N-2)\, d_{12} + \sum_{\substack{j=1 \\ j \neq i}}^{N} d_{ij} < (N-2)\, d_{1i} + \sum_{\substack{j=1 \\ j \neq 2}}^{N} d_{2j}$$

sum of all diss to i: $d_{1i} + d_{2i} + \ldots + d_{Ni}$

sum of all diss to 2: $d_{21} + d_{22} + d_{23} + \ldots + d_{2N}$

Define: $R_i = \sum_{j=1}^{N} d_{ij}$ = all distances from i.

We have with this notation

$$(N-2)d_{12} + R_i < (N-2)d_{1i} + R_2$$

Subtract $R_1 + R_2 + R_i$

$$\Rightarrow \underbrace{(N-2)d_{12} - R_1 - R_2}_{M_{12}} < \underbrace{(N-2)d_{1i} - R_1 - R_i}_{M_{1i}}$$

Define $M_{\alpha\beta}$

Defn: Let $M_{ij} = (N-2)d_{ij} - R_i - R_j$

We have shown that when 1,2 form a cherry

$$M_{12} < M_{1i} \quad \text{for any } i = 3, 4, \ldots, N$$

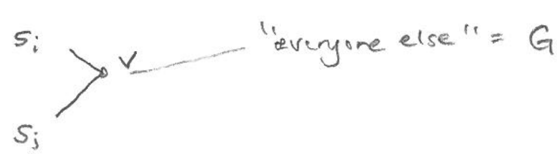
This gives the Neighbor-Joining joining criterion:

Step 1: For all pairs i,j, $i \neq j$, compute M_{ij}

Join the taxa S_i, S_j with M_{ij} smallest. (Break ties arbitrarily.)

≡ Selection criterion!

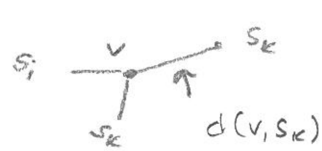
To join S_i, S_j, use the 3-point formula and FM method



save

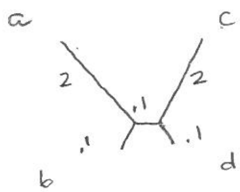
Step 2: Collapse distance table

We will replace S_i, S_j with v to get a new distance table with <u>one</u> less taxon. Get the distances $d(v, S_k)$ $k \neq i,j$ by using the 3-point formula on S_i, S_j, S_k



REPEAT until only 3 groups remain in table. Use 3-point formula to finish ⤳ the Neighbor-Joining tree.

Example: (Also illustrates that NJ recovers true tree from tree metric dissimilarity data.)



unknown T

	a	b	c	d
a		2.1	4.1	2.2
b			2.2	.3
c				2.1

Known data

$R_a = 8.4$ $\quad$ $R_c = 8.4$ $\quad$ $N-2=2$

$R_b = 4.6$ $\quad$ $R_d = 4.6$

$M_{ab} = 2d_{ab} - R_a - R_b = 2(2.1) - 8.4 - 4.6 = -8.8$ *

$M_{ac} = 2d_{ac} - R_a - R_c = 2(4.1) - 8.4 \;\; 8.4 = -8.6$

$M_{ad} = 2d_{ad} - R_a - R_d = 2(2.2) - 8.4 - 4.6 = -8.6$

* = minima!

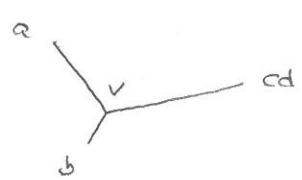
$M_{bc} = 2d_{bc} - R_b - R_d = 2(2.2) - 4.6 - 4.6 = -4.8$

$M_{bd} = 2d_{bd} - R_b - R_d = 2(.3) - 4.6 - 4.6 = -8.6$

$M_{cd} = 2d_{cd} - R_c - R_d = 2(2.1) - 8.4 - 4.6 = -8.8$ *

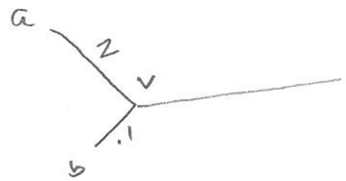
Join a, b:

	a	b	cd
a		2.1	3.15
b			1.25



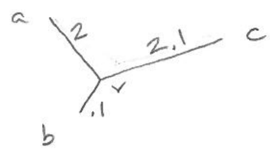
$x = d(a,v) = \frac{1}{2}\left(d(a,cd) + d(a,b) - d(b,cd)\right)$

$= \frac{1}{2}(3.15 + 2.1 - 1.25) = \frac{4}{2} = 2!$



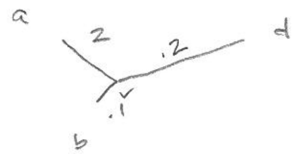
Step 2: Compute $d(v, c)$, $d(v,d)$

	a	b	c
a		2.1	4.1
b			2.2



$d(v,c) = 2.1$

	a	b	d
a		2.1	2.2
b			.3

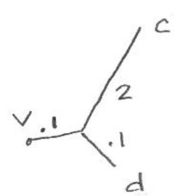


$d(v,d) = .2$

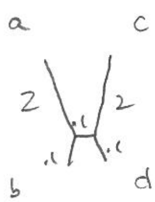
New Table:

	v	c	d
v		2.1	.2
c			2.1

only 3 groups! ⇒ end in sight



Attach at v:



NJ tree = true tree!