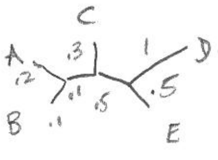
Chapter 5: DISTANCE METHODS

We have already seen metric trees and, more formally,

TREE METRICS ω

Review: $\omega : X \times X \rightarrow \mathbb{R}^{\geq 0}$ is a tree metric if there exists a tree T with exactly the pairwise distances given by ω.

Eg. Behind the scenes:



(T, ω)

The TREE METRIC ω is

	A	B	C	D	E	
A		.3	.6	1.8	1.3	etc.
B						
C						
D						

↑

This table is called a "distance table," but

CAUTION: "distance" table is used ambiguously

- when the table corresponds to a tree (T, ω)

  ⇝ a tree metric

- when the table does not correspond to a tree metric

  i.e. the table does <u>not</u> fit a tree

  for example, when pairwise numerical comparisons are computed from sequence data, or in the presence of error even if a tree metric underlies...

We will go along with this ambiguous use ... a bit. (common)

However, if a table does not come from a tree metric, (i.e. there is <u>no</u> tree with those pairwise weights) the correct term is DISSIMILARITY measure, DISSIMILARITY map, DISSIMILARITY TABLE

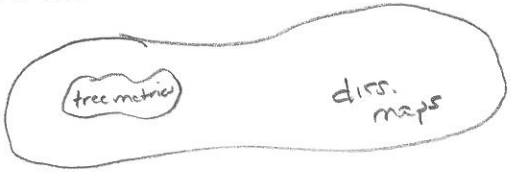
Defn: If X is a set of taxa labels, then a DISSIMILARITY MAP

is a function $\delta : X \times X \longrightarrow \mathbb{R}$ ($\mathbb{R}^{\geq 0}$ in our application)

such that
$$\begin{cases} \delta(x,x) = 0 & \text{for all } x \in X \\ \delta(x,y) = \delta(y,x) & \text{for all } (x,y) \text{ pairs } \in X \times X \end{cases}$$

Informally, a function that assigns nonnegative numbers to pairs of taxa (a,b)

Note: A tree metric ω (or d) is a dissimilarity map, but not vice versa



Given 2 sequences for taxa a, b, a natural dissimilarity map

$\delta(a,b)$ = average number of differences between sequences

a: AATCG
b: AACCG

$\delta(a,b) = 1/5$

This is called the

HAMMING DISTANCE

p-distance

uncorrected distance

uncorrected p-distance.

Note: the incorrect, but common use of distance

Distance Methods: Methods to fit dissimilarity matrix to a tree.

Caution: We will use $d(A,B)$ for distances computed from taxa A,B.

Technically, $d(A,B)$ is a <u>dissimilarity</u> between A,B.

Method 1: UPGMA ≡ Unweighted Pair Group Method with Arithmetic Mean

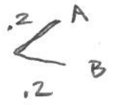
Ex. Original dissimilarity table: $n=5$ taxa

	B	C	D	E
A	.4	.5	.6	.7
B		1.1	1.9	1.6
C			1.6	1.0
D				.9

Keep this! Will be used in each step.

1) Choose smallest distance in current dis. table.

$d(A,B) = .4$ Join the two taxa A,B, placed equidistant from vertex



2) Join taxa together in agglamerate taxon, compute new current distance by averaging distances to group IN THE ORIGINAL TABLE

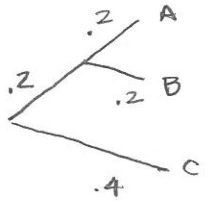
	C	D	E
AB	.8	1.25	1.15
C		1.6	1.0
D			.9

$$d(AB,C) = \frac{d(A,C) + d(B,C)}{2} = \frac{.5+1.1}{2} = .8$$

$$d(AB,D) = \frac{.6+1.9}{2} = 1.25$$

$$d(AB,E = \frac{.7+1.6}{2} = \frac{2.3}{2} = 1.15$$

$d(AB,C) = .8$ smallest!



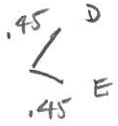
New distance table

	D	E
ABC	$1.3\overline{6}$	1.1
D		.9

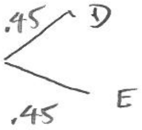
$$d(ABC,D) = \frac{d(A,D) + d(B,D) + d(C,D)}{3}$$

$$= \frac{.6 + 1.9 + 1.6}{3} = \frac{4.1}{3} = 1.3\overline{6}$$

$d(D,E)$ smallest!



$$d(ABC,E) = \frac{.7 + 1.0 + 1.6}{3} = \frac{3.3}{3} = 1.1$$



Lastly, compute FROM ORIGINAL TABLE

$$d(ABC,DE) = \frac{d(A,D) + d(B,D) + d(C,D) + d(A,E) + d(B,E) + d(C,E)}{6}$$

$$= \frac{.6 + 1.9 + 1.6 + .7 + 1.6 + 1.0}{6} = 1.2\overline{3}$$

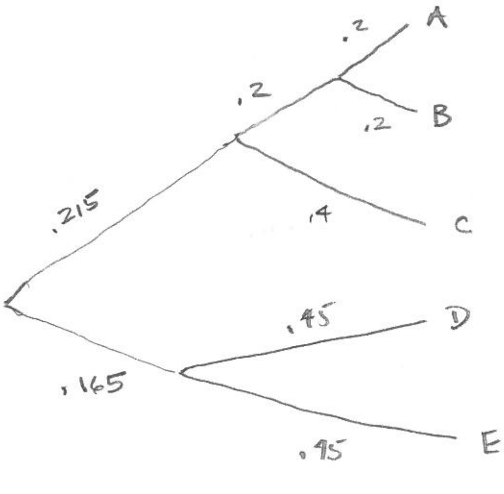
$.615 - .45 =$

Root-to-tip distance will be $\frac{1}{2}(1.23) \approx .615$

UPGMA tree:

is ROOTED, ULTRAMETRIC

binary tree.


.2
A
.2
B
.2
.215
.4
C
.45
D
.165
.45
E