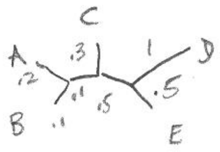
Chapter 5: DISTANCE METHODS

We have already seen metric trees and, more formally,

TREE METRICS ω

Review: $\omega : X \times X \to \mathbb{R}^{\geq 0}$ 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)

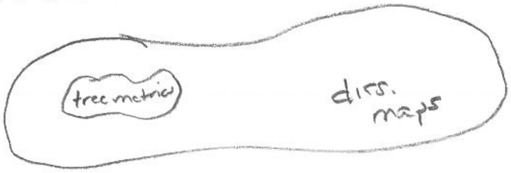
However, if a table does not come from a tree metric, (i.e. there is <u>no</u> tree with those pairwise weights) the correct term is DISSIMILARITY measure, DISSIMILARITY map, DISSIMILARITY TABLE

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 w (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

Note: the incorrect, but common use of distance

p-distance

uncorrected distance

uncorrected p-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

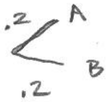
Ex. Original dissimilarity table: $n=5$ taxa

	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 **IN THE ORIGINAL TABLE** averaging distances to group

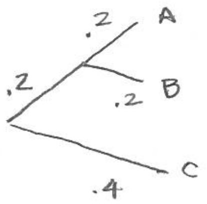
	C	D	E
AB	.8	1.25	1.15
C		1.6	1.0
D			.9

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

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

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

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



New distance table

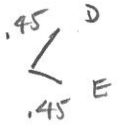
	D	E
ABC	$1.3\bar{6}$	1.1
D		.9

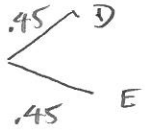
$$d(ABC,D) = \frac{d(A,D) + d(B,D) + d(C,D)}{3}$$

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

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

$d(D,E)$ smallest!





Lastly, compute FROM ORIGINAL TABLE

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

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

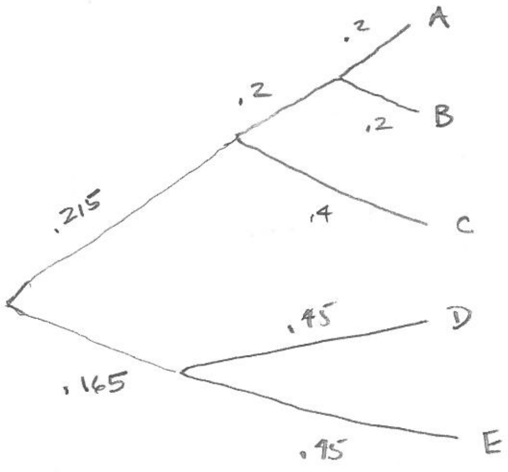
$.615 - .45 =$

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

UPGMA tree:

is ROOTED, ULTRAMETRIC

binary tree.



Example: ORIGINAL Distance Table

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

Iteration 1:

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

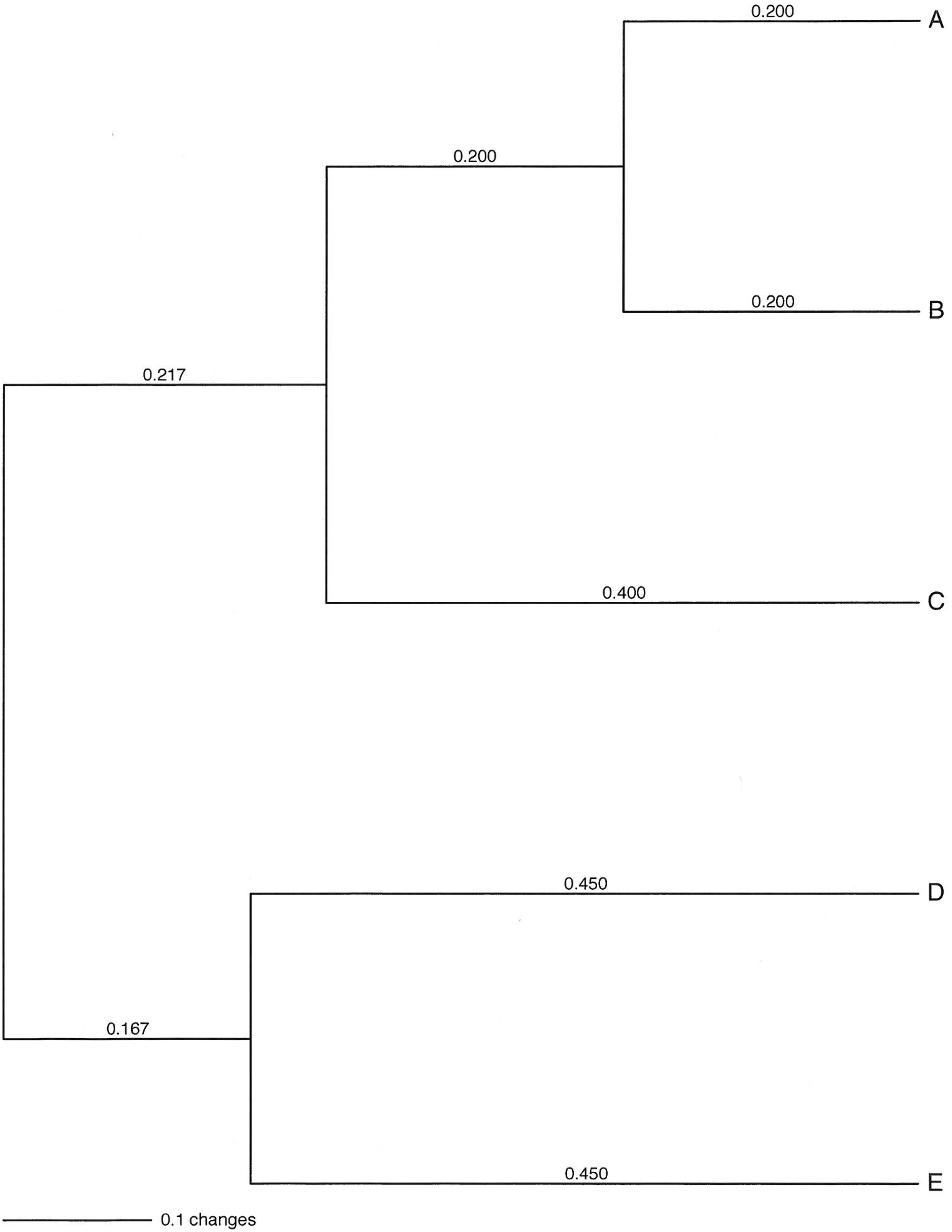
Iteration 2:

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

Iteration 3:

$d(ABC, DE) = 1.2\overline{3}$

UPGMA tree



Conclusions:

UPGMA:

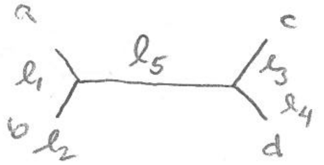
- constructs a rooted, ultrametric tree (binary)
- is fast! Instead of searching over $(2n-3)!!$ trees like parsimony, UPGMA quickly constructs a tree from dissimilarity data
- could be reasonable if one assumes/believes a molecular clock is at work.

Before continuing, review solving m equations in n unknowns ...

Suppose the taxa under study are X, with $|X| = n$. Then any dissimilarity table has $\binom{n}{2} = \frac{n(n-1)}{2}$ dissimilarities, but an (unrooted) binary tree has only $2n-3$ edges with lengths l_i, $i = 1, \dots, 2n-3$

n	# pairwise diss.	# edge length l_i
2	1	1
3	3	3
4	6	5
	⋮	⋮
10	45	17

$\binom{n}{2} \gg 2n-3$ as $n \to \infty$



Viewing the l_i as unknowns we have

$n=10$: 45 equations in 17 unknowns

$n=4$: 5 equations in 6 unknowns

$\Rightarrow$ overdetermined system (more equations than unknowns)

$\rightsquigarrow$ likely inconsistent (i.e. <u>no</u> solution)

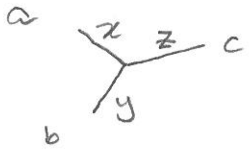
$\rightsquigarrow$ distance tables are usually <u>not</u> from tree metrics.

However, when $n=3$ there are 3 equations in 3 unknowns and the system has a solution.

Eg.

	a	b	c
a		4	3
b			5

Single unrooted tree: with $l_i = a, b, c$.



E1: $d(a,b) = 4 = x + y$

E2: $d(a,c) = 3 = x + z$

E3: $d(b,c) = 5 = y + z$

3 linear equations in 3 unknowns ⇝ consistent

Solve using linear algebra <u>OR</u> common sense.

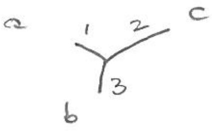
$$x = \frac{d(a,b) + d(a,c) - d(b,c)}{2} = \frac{4+3-5}{2} = 1 \qquad \boxed{x=1}$$

$$\boxed{y=3}$$

$$y = d(a,b) - x = 4-1 = 3$$

$$\boxed{z=2}$$

$$z = d(a,c) - x = 3-1 = 2$$



We use this idea → 3 pairwise distances exactly fit a tree ← to get a new distance method that does not produce ultrametric trees.

Fitch-Margoliash (FM) : means to end = NJ = neighbor-joining

Distance Table: (delete A momentarily)

	B	C	D	E
B		1.1	1.9	1.6
C			1.6	1.0
D				.9

Step 1: Choose closest taxa to join D, E

→ Change from UPGMA will use 3-point formula to join D, E

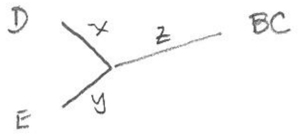
To do this, create temporary distance table

with D, E, G = everyone else = $\{B, C\}$

	D	E	BC
D		.9	1.75
E			1.3

$d(D, BC) = \frac{1.9 + 1.6}{2} = \frac{3.5}{2} = 1.75$

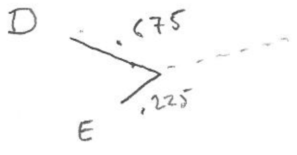
$d(E, BC) = \frac{1.6 + 1.0}{2} = \frac{2.6}{2} = 1.3$



Join chosen taxa (D, E) using 3-point formula

$$x = \frac{d(D, BC) + d(D, E) - d(E, BC)}{2} = \frac{1.75 + .9 - 1.3}{2} = .675$$

$$y = d(D, E) - x = .9 - .675 = .225$$



Keep only edges leading to D, E

Step 2: Collapse joined taxa into new group and make collapsed new distance table ⇝ identical to UPGMA

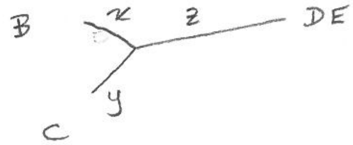
	DE	B	C
DE		1.75	1.3
B			1.1

$d(B,DE) = \frac{1}{2}(1.9+1.6) = \frac{3.5}{2} = 1.75$

$d(C,DE) = \frac{1}{2}(1.6+1.0) = 1.3$

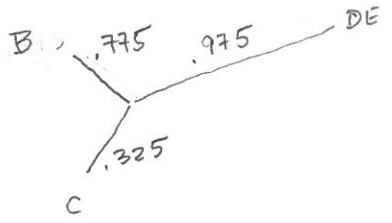
REPEAT ...

BC smallest ... only 3-groups ⇒ use 3-point formula



$x = \frac{1}{2}\left(d(B,C) + d(B,DE) - d(C,DE)\right)$

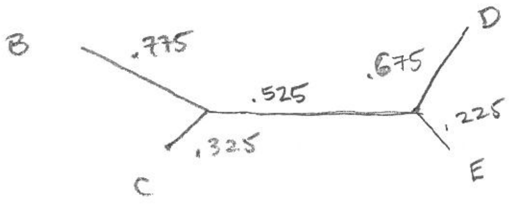
$= \frac{1}{2}\left[1.1 + 1.75 + 1.3\right] = \frac{1.55}{2} = .775$



$y = d(B,C) - .775 = 1.1 - .775 =$

$z = d(C,DE) - .325 = .975$

LAST STEP: Combine again using average



middle edge =

$.975 -$ (average dist. to DE)

$= .975 - .45$

FITCH - MARGOLIASH:

* not used in practice, but a means to ... NJ
=neighbor joining

Same distance table used for UPGMA.

Step 1: Choose closest 2 taxa to join A,B

Different!

→ Collapse all other remaing taxa into group G = {C, D, E}

and form temporary distance table for A, B, G

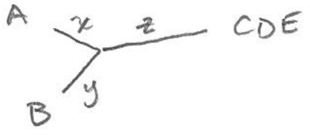
	B	CDE
A	.4	.6
B		1.8

$$d(A, CDE) = \frac{d(A,E) + d(A,D) + d(A,C)}{3}$$

$$= \frac{.5 + .6 + .7}{3} = \frac{1.8}{3} = .6$$

$$d(B, CDE) = \frac{1.1 + 1.9 + 1.6}{3} = \frac{4.6}{3} = 1.8$$

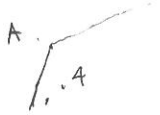
Use 3-point formula to fit chosen taxa (A,B) to tree



$$\frac{1}{2}\left(d(A,CDE) + d(A,B) - d(B,CDE)\right) = x$$

$$\frac{1}{2}\left[.6 + .4 - 1.8\right] < 0$$

Reset to zero!



Toss away temporary table

Step 2: Join AB into group for rest of analysis, Create new current distance table from original table just as in UPGMA.

REPEAT ...