

28 Jan 2015

①

① biologists relied on morphological features — phenotypes such as beak shapes / presence — absence of fins to construct evolutionary trees.

but today use DNA sequences for estimating ex. ET sequences

you take a gene and compare ~~the~~ it across species.

out of Africa hypothesis: mitochondrial DNA mtDNA
humans have a common ancestor who lived in Africa
200k yrs ago.

mtDNA of different races, personalities, nationalities

mtDNA passes from mother to children w/ recombining
with father's DNA

it quickly accumulated mutations

quick-ticking molecular clock.

Africa

Asia

Australia

Caucasian

New Guinea

tree shows what? trunk splits into two branches

one branch — only africans

other branch — modern africans &
everyone else.

→ first modern humans form the trunk & the
first branch of the tree

→ second branch a subgroup that left Africa & spread to
rest of the world.

our species is relatively young

African samples had most mutations $\Rightarrow$ African lineage is the oldest & modern humans trace their roots back to Africa.

modern man emerged from Africa $\approx$ 200k yrs ago

lots of controversy.

(2^o) tree estimation

tree reconstruction - error free process leading to the correct tree

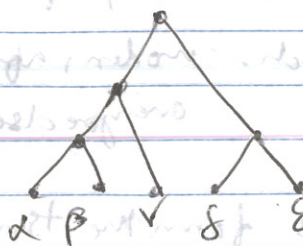
BT - binary - all internal nodes connect to 1 or 3 nodes
root, at most one root, connects with 2.

tree - no root $\Rightarrow$ unrooted tree

what kind of species should we compare?

natural to construct a tree relating various kinases
various hemoglobins

eg.



rooted tree

5 species

length of an edge = evolutionary time

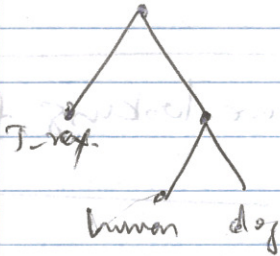
↑ length(edge) = longer the evolutionary time
separating the nodes at end of edge

why are $\alpha, \beta, \gamma, \delta, \epsilon$ at the same level?

vertical height?

because they are the existing species.

internal $\Rightarrow$ hypothetical ancestral species (extinct)



unrooted tree

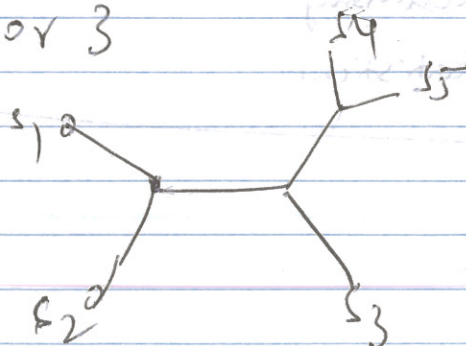
what about UT (unrooted tree)?

edges depict evolution relationships
but do not indicate a direction of evolution
which evolutionary time
flowed between internal nodes.

how many edges for each node in an unrooted tree?

1 or 3

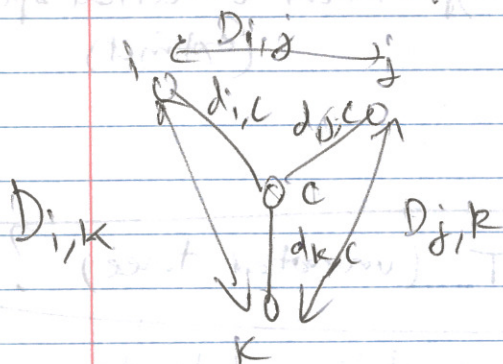
eg:



Q: Distance-based phylogeny problem

Simple version: given a weighted tree T with n leaves, we can compute the length of the path between any two leaves $i, j \rightarrow d_{ij}(T)$

difficult version: given a $n \times n$ matrix of ^{evolutionary} distances $D(i, j)$ is there a tree T with n leaves that fits the data, that is $d_{ij}(T) = D_{ij}$



What are we looking for?

- Shape of a tree T
- that fits the distance matrix
- weights for each edge in tree

tree
how many types of reconstruction methods

- distance
- parsimony
- statistical

3

Distances

Ideal definition: # of yrs since their most common ancestor

Use of Surrogates: surrogate distances are used

for now assume distances are known in years.

distance as a metric

in euclidean space $x, y \in \mathbb{R}^n$

$$1 \quad d(x, y) \geq 0$$

$$2 \quad d(x, y) = 0 \Rightarrow x = y$$

$$3 \quad d(x, y) = d(y, x)$$

$$4 \quad d(x, y) \leq d(x, z) + d(z, y)$$

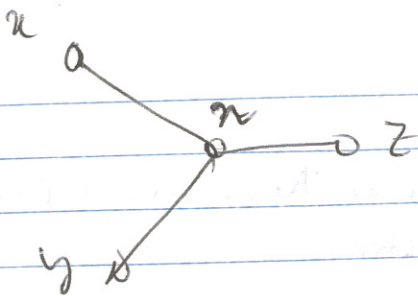
for any 2 elements

distance tree-derived

S = set of species

you have a tree with these species at the leaves st distance $d(x, y)$ between any pair is the sum of the lengths of the edges joining them.

what properties does the tree-derived distance satisfy



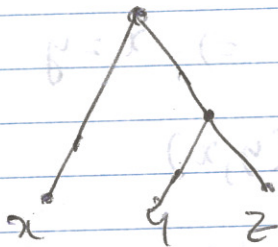
distances satisfy strict inequality

$$d(x, z) < d(x, y) + d(y, z)$$

strict inequality — must hold

$\forall x, y, z$ in the tree

rooted trees have additional constraints



$$d(x, y) = d(x, z)$$

— 2 distances are same

$$d(y, z) < d(x, y)$$

$$d(y, z) < d(x, z)$$

& greater than third

$d \Rightarrow$ ultrametric

$\Rightarrow$ tree-derived distance measure on a rooted tree of extant species with exact edge lengths is ultrametric

$\Rightarrow$ converse is true

$\Rightarrow$ how to deduce such a tree = ? unique except for trivial changes

Also

$\Rightarrow$ not all tree-derived distance measures are ultrametric but we can recover a unique unrooted tree!

④

Rooted Tree - ultrametric case

④

Input: $\{s_1, s_2, \dots, s_n\} \subseteq S$
 $d(a, b) \neq d(a, c) \text{ in } S$

Proof by induction:

Step 1 — ~~show for two species~~

construct the ~~tree~~ tree for two species s_1 show it is unique

2 — show that if we can reconstruct

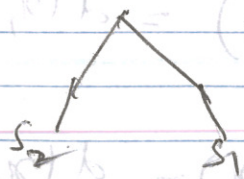
the tree correctly and uniquely for n species

$s_1, s_2, \dots, s_m$, for some $m \geq 2$

we can do this also for $m+1$ species

$s_1, s_2, \dots, s_{m+1}$.

both s_1 & s_2 have to be
 $d(s_1, s_2)/2$ $d(s_1, s_2)/2$ equidistant from root.
 clearly there is a unique soln



trivial change. ignore

Now, suppose there are $m+1$ species, $m \geq 2$.

By induction, assume we have a unique tree reconstructed for $S_1, S_2, \dots, S_m$.

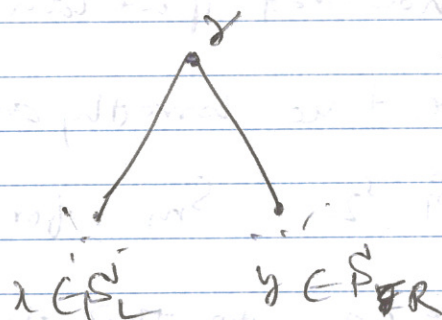
The species $S_1, S_2, \dots, S_m$ divide into two sets

S_L : species down the left edge from the root

S_R : species right edge

$x \in S_L$

$y \in S_R$



S_{m+1}

$d(S_{m+1}, x)$

$d(S_{m+1}, y)$

$d(x, y)$

} 3 distances

Ultrametric possibilities ① $d(S_{m+1}, x) = d(S_{m+1}, y)$

② $d(S_{m+1}, y) = d(x, y)$

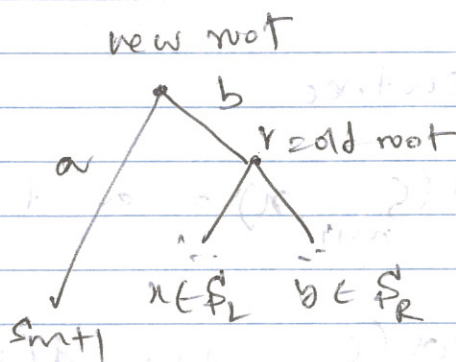
③ $d(S_{m+1}, x) = d(x, y)$

say ① $d(S_{m+1}, x) = \frac{d(S_{m+1}, y) + d(x, y)}{2}$ holds then

$$d(x, y) < d(S_{m+1}, x)$$

$$< d(S_{m+1}, y)$$

where to place S_{m+1} ?



$$a = \frac{d(S_{m+1}, x) + d(S_{m+1}, y)}{2}$$

$$b = \frac{d(S_{m+1}, x) + d(x, y)}{2}$$

we know

$$d(x, y) < d(S_{m+1}, x)$$

$$\Rightarrow b > 0$$

$\Rightarrow$ we have a tree that gives correct distances between S_{m+1} , x and y .

$\Rightarrow$ we must check that this placement gives correct distance from S_{m+1} to any other species

for any species z in S_L or S_R

$$d(z, \text{old root}) = d(x, y)/2$$

$$\Rightarrow d(s_{m+1}, z) \stackrel{?}{=} a + b + d(x, y)/2 \quad \text{) prove}$$

in the new tree

$$d(s_{m+1}, x) = a + b + \frac{d(x, y)}{2}$$

sufficient to show that $d(s_{m+1}, z) = d(s_{m+1}, x) \quad \forall z \in S_L \cup S_R$

Suppose $z \in S_L$. By induction hypo, the tree gives the correct distance between x & $z \Rightarrow$

$$d(x, z) < d(x, y)$$

$$d(x, z) < d(x, y) < d(s_{m+1}, x)$$

by ultrametric property

$$d(s_{m+1}, x) = d(s_{m+1}, z)$$

~~Repeat the~~ symmetric argument when $z \in S_R$.

Case ii) $d(s_{m+1}, y) = d(x, y)$

ultrametric prop says: $d(s_{m+1}, x) < d(x, y)$

since x is in S_L , y is in S_R

$$d(x, y) = d(x, z) \quad \forall z \text{ in } S_R$$

Therefore, $d(s_{m+1}, x) < d(x, z) \quad \forall z \text{ in } S_R$

and $d(s_{m+1}, z) = d(x, z) \quad \forall z \text{ in } S_R$

thus, s_{m+1} is equidistant from every member in S_R .

2 possibilities

s_{m+1} equidistant from every member γ S_L or

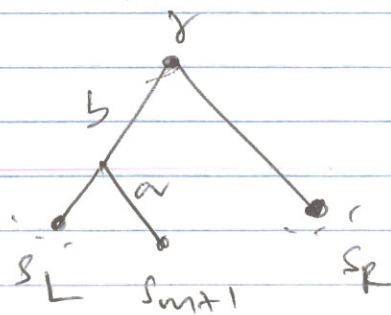
it is not.

If s_{m+1} equidistant from S_L add s_{m+1}

$$a = \frac{d(s_{m+1}, x)}{2}$$

$$b = \frac{d(s_{m+1}, y)}{2} - a$$

$b > 0$ because $d(s_{m+1}, y) > d(s_{m+1}, x)$



check that this allocation gives correct distances.

— for z in S_L show $d(s_{m+1}, z) = 2a$

$$\text{i.e. } d(s_{m+1}, z) = d(s_{m+1}, x)$$

which is true because we assumed that s_{m+1} is equidistant from every one in S_L

— z in S_R then $d(s_{m+1}, z) = a + b + \frac{d(x, y)}{2}$

$$= \frac{1}{2} (d(s_{m+1}, y) + d(x, y))$$

$$= \frac{1}{2} (d(x, y) + d(x, y))$$

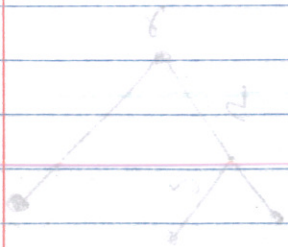
$$= d(s_{m+1}, z)$$

finally:

s_{m+1} is equidistant from every member of S_R

but is not equidistant from member of S_L

lets look at the subtrees T_L and T_R



T_L tree has $m-1$ or fewer leaves so we know

S_{m+1} can be uniquely added to this tree. &

all distances between S_{m+1} and rest of the leaves

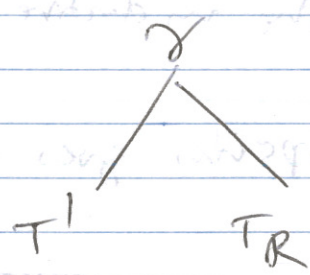
T_L are correct! call this new tree T' !

Since S is NOT equidistant to every leaf in

T_L , T' has the same root as T_L .

therefore we can put back together the whole tree

relating $S_1, S_2, \dots, S_{m+1}$ as



we now show that S_{m+1} has the correct distances

to the rest of the species $S_1, \dots, S_m$.

Since we know T' has been correctly reconstructed, this

is true for all species in S_L .

Suppose z is in S_R .

Now x is in S_L , so the tree

constructs the distance from S_{m+1} to z to be $d(z, x)$

therefore show that $d(x, z) = d(S_{m+1}, z)$

we know $d(x, y) = d(x, z)$

therefore $d(x, S_{m+1}) < d(x, z)$

ultrametric property gives $d(x, z) = d(S_{m+1}, z)$

proof is constructive in that it provides an
also for reconstructing the tree.

You add species to the tree one by one in no
particular order, with the inductive step showing
explicitly where each species goes next.

UPGMA (Sokal & Mittermeier 1958)

unweighted pair group method using arithmetic averages

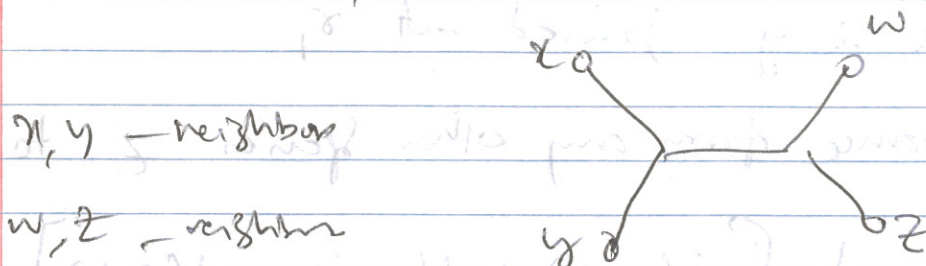
feed ultrametric distances $\Rightarrow$ gives you a tree as just
outlined

but applicable to any set of distances

If not ultrametric, it cannot reconstruct a tree that
give back the distance.

UPGMA

2 species are neighbors if the path between them contains only one node.



for a tree whose derived distance satisfies

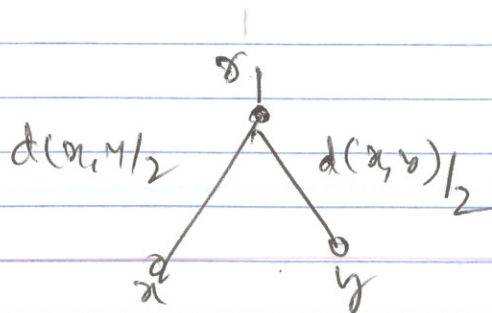
UM property, pairs of species that are closest together wrt distance matrix are neighbors

UPGMA

$d(i, j)$ dist between i & j species.

define groups G_u G_v

$$d(G_u, G_v) = \frac{1}{2} \sum_{\substack{x, y \\ x \in G_u \\ y \in G_v}} d(x, y)$$



assumed with a defn of

distance between groups

reversibly build a tree from bottom-up.

this procedure gives tree topology.

then solve for correct lengths of edges in the tree.

first step x & y joined at x_1

then distance from any other species z to

$$x_1 = \frac{1}{2} [d(x, z) + d(y, z) - d(x, y)]$$

