Week 5 Phylogenomics

Friday, 18 February 2022 8:02 AM

African ancestry (American SW)

20x coverage LCT ; rs4988235 SNP

70% G (ref) ; 30% A (alternate)

① Compute genotype likelihoods.

② which is the most likely genotype at this locus?

n = total # of reads (20)

TRUTH | GENOTYPE LIKELIHOOD

GG | ${}^{n}C_{k}\, \epsilon^{k} (1-\epsilon)^{n-k}$ $k = 14$ $= P(D|GG)$

AA | ${}^{n}C_{n-k}\, \epsilon^{n-k} (1-\epsilon)^{k} = P(D|AA)$

AG | ${}^{n}C_{k} \left(\frac{1}{2^n}\right) = P(D|AG)$

① $P(D|GG) = {}^{20}C_{6} \left(\frac{6}{20}\right)^{6} \left(\frac{14}{20}\right)^{14} = 0.191639$

② $P(D|AA) = {}^{20}C_{14} \left(\frac{14}{20}\right)^{6} \left(\frac{6}{20}\right)^{14} = 0.002181$

③ $P(D|AG) = {}^{n}C_{k} \left(\frac{1}{2^n}\right) = {}^{20}C_{14} \left(\frac{1}{2^{20}}\right) = 0.0369$

$P(GG|D) \propto P(D|GG) \times P(GG)$

$\propto 0.191639 \times 0.705 = 0.135$

$P(AA|D) \propto P(D|AA) \times P(AA)$

$\propto 0.002181 \times 0.049 = 0.000106$

$P(AG|D) \propto P(D|AG) \times P(AG)$

$= 0.009074$

$\propto$ 0.0369 x

TYPES Mutations

PURINES – A, G
PYRIMIDINES – C, U, T

SUBSTITUTION (SNPs)
- transitions (PUR ↔ PUR, PYR ↔ PYR)] transitions more likely
- transversions (PUR ↔ PYR)]

synonymous vs non-synonymous] synonymous more likely

TYPES INDELS] indels are less likely than subs.

MSA → sequence similarity
- LESS → some degree of sequence similarity
  - NO → LIKELIHOOD – BASED (Bayesian)
  - YES → DISTANCE – BASED (NJ, UPGMA)
- LOTS → PARSIMONY

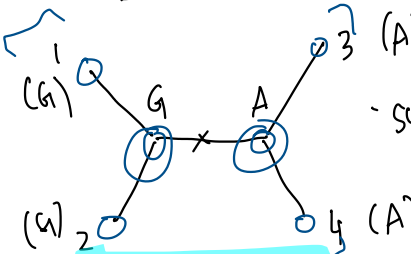
PARSIMONY – MINIMUM EVOLUTION

	1	2	3	4
INDIV 1	A	G	T	A
2	C	G	T	G
3	T	A	T	G
4	G	A	T	G

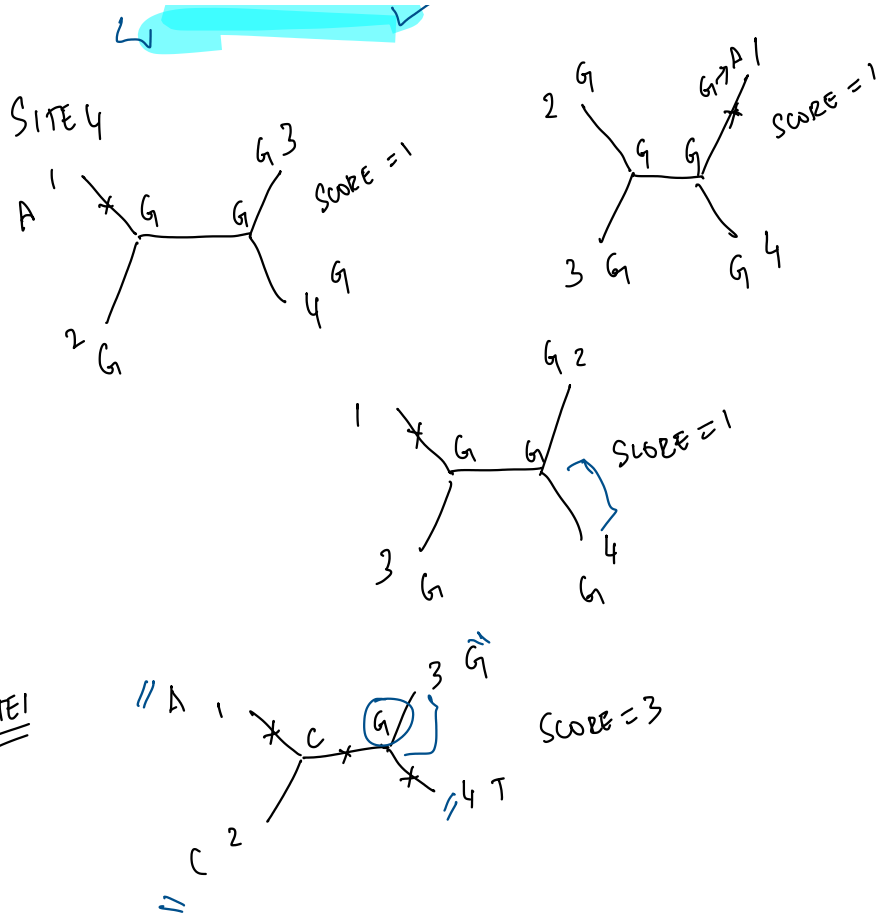
SITE 3 – invariant

SITES 1, 2, 4 – SEGREGATING

SITE 2



- score = 1]



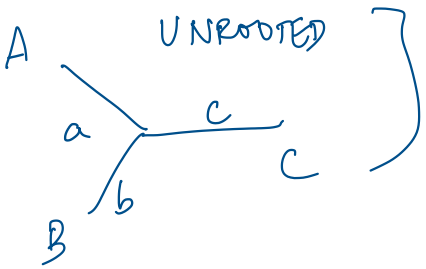
	1	2	3	4
1	0	2	3	3
2		0	2	2
3			0	1
4				0

FITCH & MARGOLIASH

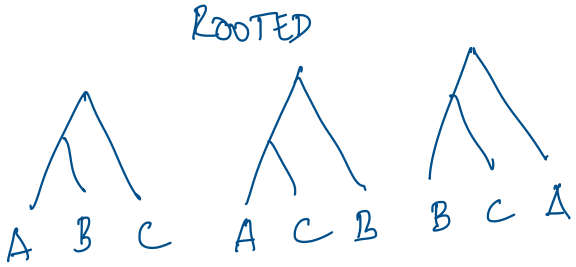
(1) MOLECULAR CLOCK → mutations accumulate at a constant rate along a branch

(2) INFINITE SITES MODEL → a mutation occurs just once at a site]

	A	B	C
A	-	22	39
B		-	41
C			-



NEIGHBOR - JOINING



(1) Always start tree with "STAR" TOPOLOGY with 3 branches

(2)

$$a + b = 22 \quad -(1)$$

$$a + c = 39 \quad -(2)$$

$$b + c = 41 \quad -(3)$$

Subtract (2) from (1)

$$a + b - a - c = 22 - 39$$

$$b - c = -17 \quad -(4)$$

add (3) and (4)

$$b + c + b - c = 41 - 17$$

$$2b = 24$$

$b = 12$

$\Rightarrow$ put this in (3)

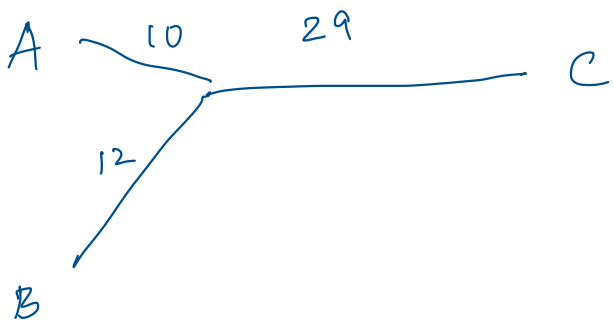
$12 + c = 41$

$\Rightarrow c = 29$

put $b = 12$ in (1)

$\Rightarrow \quad a + 12 = 22$

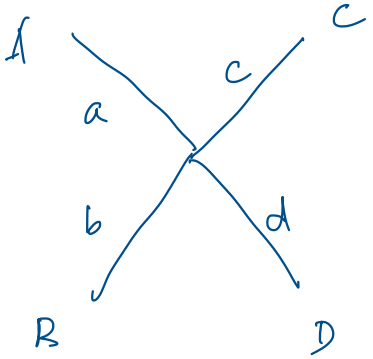
$\Rightarrow a = 10$



---

eg.

	A	B	C	D
A		2	3	3 //
B			2	2
C				1
D				



NJ STEP

C 0.5 C $\quad x = 2$ (A B)

D

	C	D	(A,B)
C		1	$\frac{AC+BC}{2} = \frac{3+2}{2} = 2.5$
D			$\frac{AD+BD}{2} = \frac{3+2}{2} = 2.5$
(A,B)			

$c + x = 2.5$ — (1)

$d + x = 2.5$ — (2)

$c + d = 1$ — (3)

(1) − (2) : $c + x - d - x = 0$

$\Rightarrow c - d = 0$

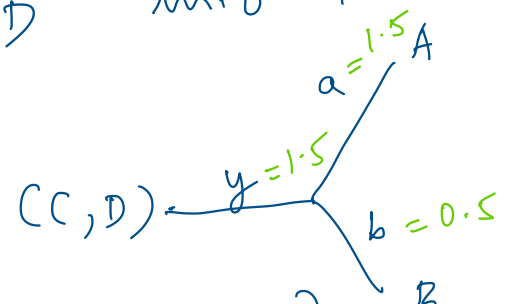
$c + d = 1$

$2c = 1$

$\Rightarrow c = 0.5$

$\therefore d = 0.5$

Join C, D into 1 node



	A	B	(C,D)
			AC + AD / 2 = 3+3 / 2 = 3

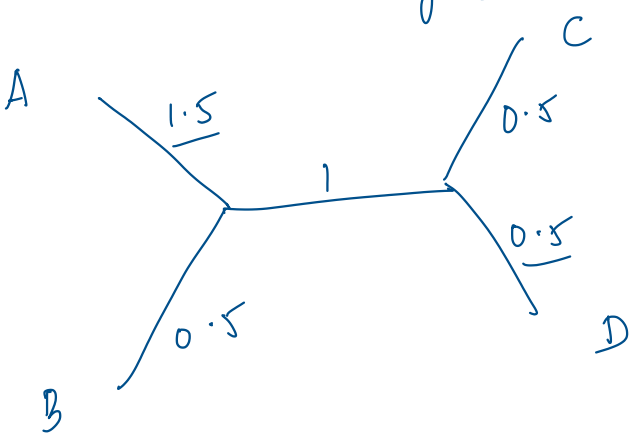
A		2	$\frac{AC+AD}{2} = \frac{\quad}{2}$
B			$\frac{BC+BD}{2} = \frac{2+2}{2} = 2$
(C, D)			

$$a + b = 2 \quad -(1)$$
$$a + y = 3 \quad -(2)$$
$$b + y = 2 \quad -(3)$$

(1) − (2):

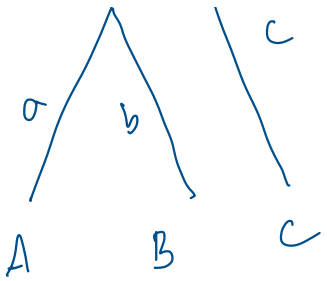
$$a + b - a - y = 2 - 3$$
$$b - y = -1$$
$$b + y = 2$$
$$2b = 1$$
$$\Rightarrow b = 0.5$$

$$a = 1.5$$
$$y = 1.5$$



$a = b$

UPGMA



$a + x = c$

ULTRAMETRIC TREES

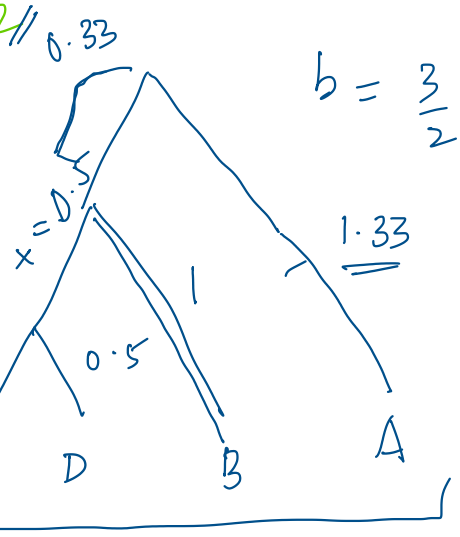
	A	B	C	D
A		2	3	3
B			2	2
C				1
D				

$c + d = 1$

$c = d$

$\Rightarrow c = d = 0.5$

	(C,D)	A	B
(C,D)		AC+AD / 2 = 3	BC+BD / 2 = 2
A			2
B			



$b = \frac{3}{2} = 1.5$

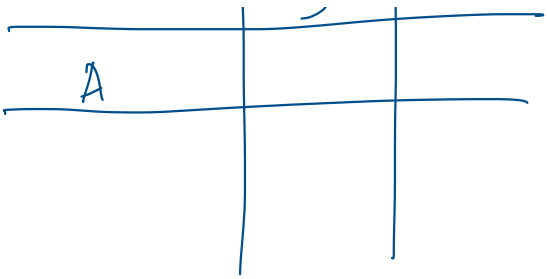
UPGMA

$x + 0.5 = b$

	A	(B,C,D)
(B,C,D)	AB+AC+AD / 3	—

$$\frac{AB + AC + AD}{3}$$

$$= 2 + 3 + 3$$



$$= \frac{8}{3} = 2.66$$

Observed data = D

H_1 = TOPOLOGY 1 , H_2 = TOPOLOGY 2

① $P(H_1|D) = \dfrac{P(D|H_1) \times P(H_1)}{P(D)}$ Bayes theorem

② $P(H_2|D) = \dfrac{P(D|H_2) \times P(H_2)}{P(D)}$

① ÷ ②

$$\Rightarrow \frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1) \times P(H_1)}{P(D|H_2) \times P(H_2)}$$

Bias = $P(H) = p \Rightarrow P(T) = 1-p$

D = HHTTHTHHTTT

$L = P(D|p) = p \times p \times (1-p) \times (1-p) \times (1-p)^3$

$\times p \times (1-p) \times p \times p \times (1-p)$

$$L = p^5 (1-p)^6 \Longrightarrow$$

SOLUTION 1 →

$$\frac{dL}{dp} = 5p^4(1-p)^6 - 6p^5(1-p)^5 = 0$$

$$p^4(1-p)^5 [5(1-p) - 6p] = 0$$

$$\Rightarrow \hat{p} = 5/11$$

SOLUTION 2 →

natural log on both sides

$$\ln L = 5 \ln p + 6 \ln (1-p)$$

$$\frac{d(\ln L)}{dp} = \frac{5}{p} - \frac{6}{1-p} = 0$$

$$\Rightarrow \hat{p} = 5/11$$

ASSUME :

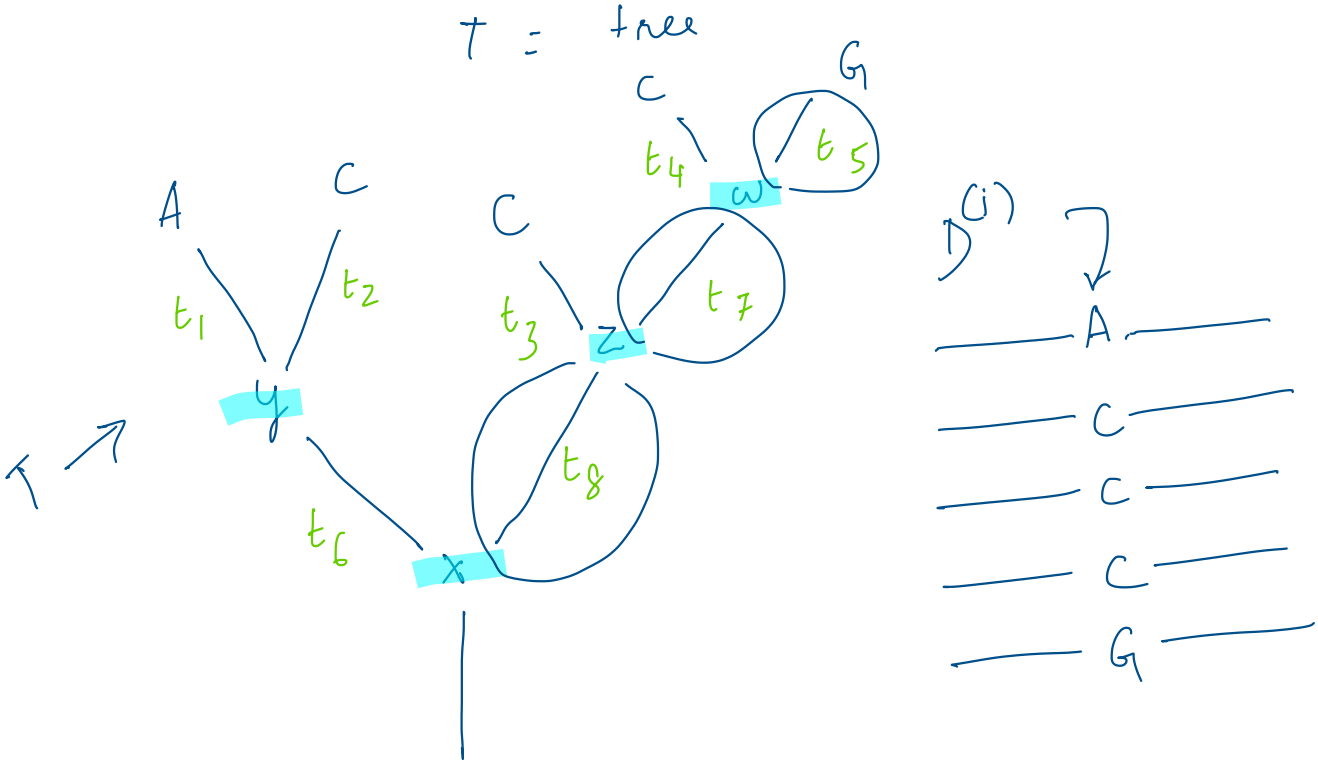
(1) evolution on different sites are independent of each other

(2) evolution on different lineages are independent of each other.

T = tree (topology) ; D = genetic data

$$\Rightarrow L = p(D|T) = \prod_{i=1}^{m} p(D^{(i)}|T) \leftarrow$$

$D^{(i)}$ = data at site i



$$p(D^{(i)}|T) = \sum_x \sum_y \sum_z \sum_w p(A,C,C,C,G,x,y,z,w|T)$$

$$x, y, z, w = \{A, C, G, T\}$$

$$= \underline{p(x)} \times \overline{p(y|x, t_6)} \times A(A|y, t_1)$$

$$\times p(c|y, t_2) \times p(z|x, t_8)$$

$$\times p(c|z, t_3) \times p(w|z, t_7)$$

$$\times p(c|w, t_4) \times p(G|w, t_5)$$