Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 20

1

TODAY'S TOPICS

- ➢ Distance estimation
  - • Nucleotide and amino acid substitution models
  - • Base composition bias, saturation
  - • Gamma distance
  - • Synonymous & nonsynonymous distances
- ➢ Phylogenetic reconstruction
  - • Terminologies
- ➢ Assignment 9

2

Distance estimation for nucleotide substitutions

➢ Jukes-Cantor (one-parameter) method

Jukes and Cantor (1969)

	A	C	G	T
A	–	α	α	α
C	α	–	α	α
G	α	α	–	α
T	α	α	α	–

All substitutions occur with equal probability

[Jukes-Cantor model of nucleotide substitutions]

$$k = -\frac{3}{4}\ln(1-\frac{4}{3}p)$$

k: Expected number of nucleotide substitutions per site or Distance

p: Proportion of nucleotide differences (observed)

$$V(k) = \frac{9p(1-p)}{(3-4p)^2 L}$$

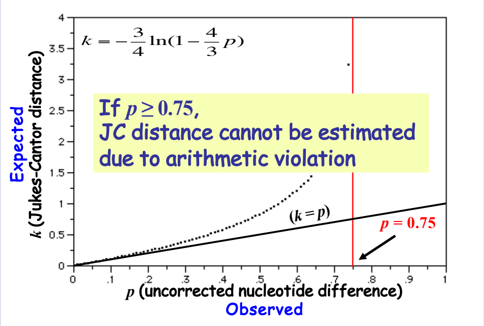
$$\sigma(k) = \frac{3}{(3-4p)}\sqrt{\frac{p(1-p)}{L}}$$

L: number of nucleotide positions compared

(Derivation of the JC equation: a note on Canvas)

3

Distance estimation for nucleotide substitutions



If $p \geq 0.75$, JC distance cannot be estimated due to arithmetic violation

4

Distance estimation for nucleotide substitutions

➢ Kimura two-parameter method Kimura (1980)

	A	C	G	T
A	–	β	α	β
C	β	–	β	α
G	α	β	–	β
T	β	α	β	–

Difference in Ts and Tv substitutions (usually Ts > Tv) can be considered

[Kimura 2-parameter model of nucleotide substitutions]

$$k = \frac{1}{2}\ln\left[\frac{1}{(1-2P-Q)}\right] + \frac{1}{4}\ln\left[\frac{1}{(1-2Q)}\right]$$

P: Proportion of transitional (Ts) differences

Q: Proportion of transversional (Tv) differences

$$V(k) = \frac{1}{L}\left[P\left\{\frac{1}{(1-2P-Q)}\right\}^2 + Q\left\{\frac{1}{(2-4P-2Q)} + \frac{1}{(2-4Q)}\right\}^2 - \left\{\frac{P}{(1-2P-Q)} + \frac{Q}{(2-4P-2Q)} + \frac{Q}{(2-4Q)}\right\}^2\right]$$

L: number of nucleotide positions compared

5

Distance estimation for nucleotide substitutions

```
ACTGTAGGAATCGC
:X::X:X:::::::
AATGCAAGAATCGC
```

Number of differences = 3

Ts = 2, Tv = 1

Alignment length = 14

- Without multiple-hit correction (p-distance):

$$p = \frac{n_d}{L} \qquad V(p) = \frac{p(1-p)}{L}$$

$p = 0.214 \pm 0.110$

- Jukes-Cantor distance:

$$k = -\frac{3}{4}\ln(1-\frac{4}{3}p) \qquad V(k) = \frac{9p(1-p)}{(3-4p)^2 L}$$

k = -3/4ln(1 - 4x0.214/3)
= 0.252 ± 0.154

- Kimura 2-parameter distance:

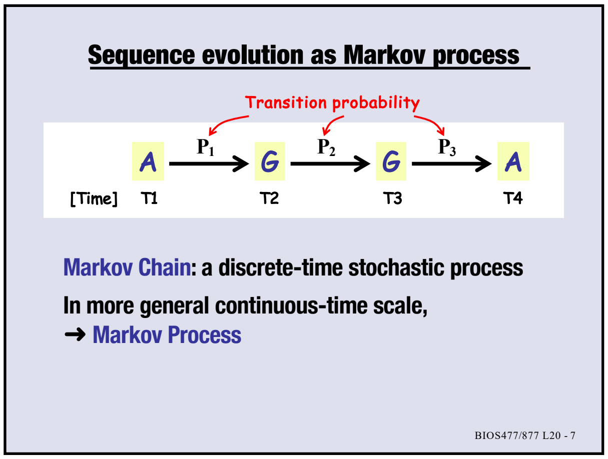
$$k = \frac{1}{2}\ln\left[\frac{1}{(1-2P-Q)}\right] + \frac{1}{4}\ln\left[\frac{1}{(1-2Q)}\right]$$

P = 2/14, Q = 1/14

k = 1/2ln[1/(1-4/14-1/14)]+1/4ln[1/(1-2/14)]
= 0.259 ± 0.143

$$V(k) = \frac{1}{L}\left[P\left\{\frac{1}{(1-2P-Q)}\right\}^2 + Q\left\{\frac{1}{(2-4P-2Q)} + \frac{1}{(2-4Q)}\right\}^2 - \left\{\frac{P}{(1-2P-Q)} + \frac{Q}{(2-4P-2Q)} + \frac{Q}{(2-4Q)}\right\}^2\right]$$

6



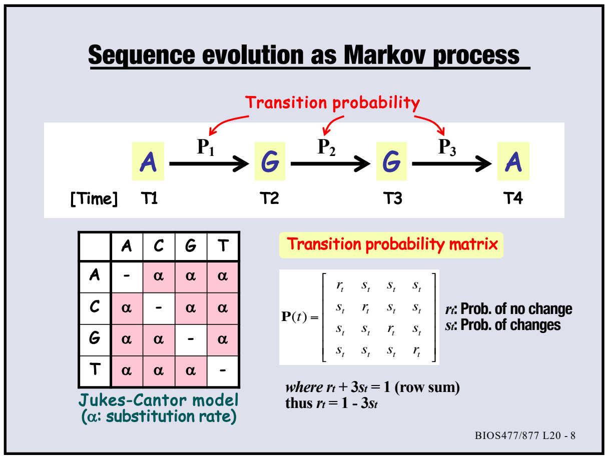
Sequence evolution as Markov process

Transition probability
A
P_1
G
P_2
G
P_3
A
[Time]
T1
T2
T3
T4

Markov Chain: a discrete-time stochastic process

In more general continuous-time scale,
➜ **Markov Process**

7



Sequence evolution as Markov process

Transition probability
A
P_1
G
P_2
G
P_3
A
[Time]
T1
T2
T3
T4

	A	C	G	T
A	-	α	α	α
C	α	-	α	α
G	α	α	-	α
T	α	α	α	-

Jukes-Cantor model
(α: substitution rate)

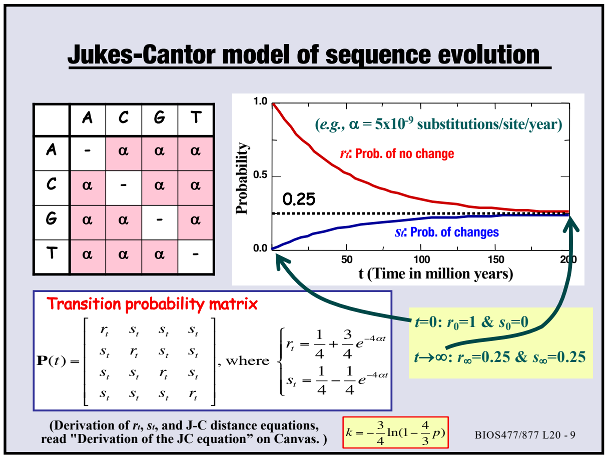
Transition probability matrix

$$\mathbf{P}(t)=\begin{bmatrix} r_t & s_t & s_t & s_t \\ s_t & r_t & s_t & s_t \\ s_t & s_t & r_t & s_t \\ s_t & s_t & s_t & r_t \end{bmatrix}$$

r_t: Prob. of no change
s_t: Prob. of changes

where $r_t + 3s_t = 1$ (row sum)
thus $r_t = 1 - 3s_t$

8



Jukes-Cantor model of sequence evolution

	A	C	G	T
A	-	α	α	α
C	α	-	α	α
G	α	α	-	α
T	α	α	α	-

Transition probability matrix

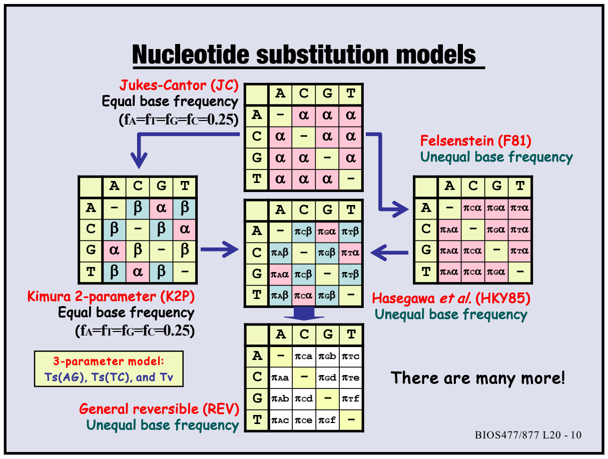
$$\mathbf{P}(t)=\begin{bmatrix} r_t & s_t & s_t & s_t \\ s_t & r_t & s_t & s_t \\ s_t & s_t & r_t & s_t \\ s_t & s_t & s_t & r_t \end{bmatrix}, \text{ where } \begin{cases} r_t = \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} \\ s_t = \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} \end{cases}$$

t=0: r_0=1 & s_0=0
$t\rightarrow\infty$: r_∞=0.25 & s_∞=0.25

(Derivation of r_t, s_t, and J-C distance equations, read "Derivation of the JC equation" on Canvas.)

$$k=-\frac{3}{4}\ln\left(1-\frac{4}{3}p\right)$$

9



Nucleotide substitution models

Jukes-Cantor (JC)
Equal base frequency
(f_A=f_T=f_G=f_C=0.25)

	A	C	G	T
A	–	α	α	α
C	α	–	α	α
G	α	α	–	α
T	α	α	α	–

Felsenstein (F81)
Unequal base frequency

	A	C	G	T
A	–	$\pi_C\alpha$	$\pi_G\alpha$	$\pi_T\alpha$
C	$\pi_A\alpha$	–	$\pi_G\alpha$	$\pi_T\alpha$
G	$\pi_A\alpha$	$\pi_C\alpha$	–	$\pi_T\alpha$
T	$\pi_A\alpha$	$\pi_C\alpha$	$\pi_G\alpha$	–

	A	C	G	T
A	–	β	α	β
C	β	–	β	α
G	α	β	–	β
T	β	α	β	–

Kimura 2-parameter (K2P)
Equal base frequency
(f_A=f_T=f_G=f_C=0.25)

	A	C	G	T
A	–	$\pi_C\beta$	$\pi_G\alpha$	$\pi_T\beta$
C	$\pi_A\beta$	–	$\pi_G\beta$	$\pi_T\alpha$
G	$\pi_A\alpha$	$\pi_C\beta$	–	$\pi_T\beta$
T	$\pi_A\beta$	$\pi_C\alpha$	$\pi_G\beta$	–

Hasegawa *et al.* (HKY85)
Unequal base frequency

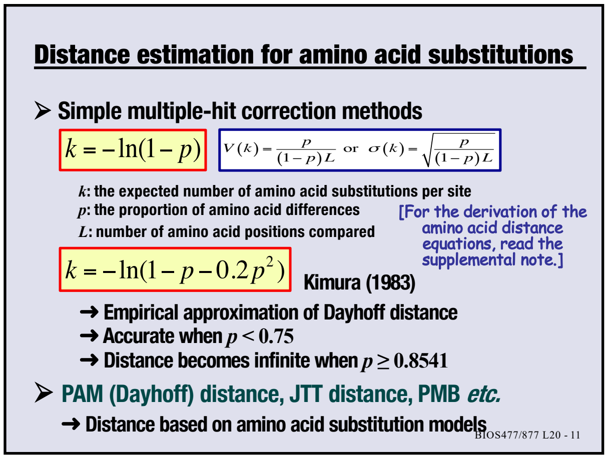
3-parameter model:
Ts(AG), Ts(TC), and Tv

	A	C	G	T
A	–	$\pi_C a$	$\pi_G b$	$\pi_T c$
C	$\pi_A a$	–	$\pi_G d$	$\pi_T e$
G	$\pi_A b$	$\pi_C d$	–	$\pi_T f$
T	$\pi_A c$	$\pi_C e$	$\pi_G f$	–

General reversible (REV)
Unequal base frequency

There are many more!

10



Distance estimation for amino acid substitutions

➢ Simple multiple-hit correction methods

$$k=-\ln(1-p)$$

$$V(k)=\frac{p}{(1-p)L} \quad \text{or} \quad \sigma(k)=\sqrt{\frac{p}{(1-p)L}}$$

k: the expected number of amino acid substitutions per site
p: the proportion of amino acid differences
L: number of amino acid positions compared

[For the derivation of the amino acid distance equations, read the supplemental note.]

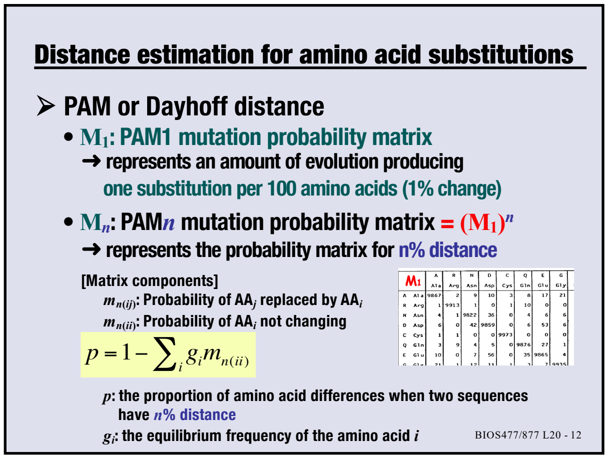
$$k=-\ln(1-p-0.2p^2)$$

Kimura (1983)

➜ Empirical approximation of Dayhoff distance
➜ Accurate when $p < 0.75$
➜ Distance becomes infinite when $p \geq 0.8541$

➢ PAM (Dayhoff) distance, JTT distance, PMB *etc.*
➜ Distance based on amino acid substitution models

11



Distance estimation for amino acid substitutions

➢ PAM or Dayhoff distance

- M_1: PAM1 mutation probability matrix
  ➜ represents an amount of evolution producing one substitution per 100 amino acids (1% change)
- M_n: PAMn mutation probability matrix = $(M_1)^n$
  ➜ represents the probability matrix for n% distance

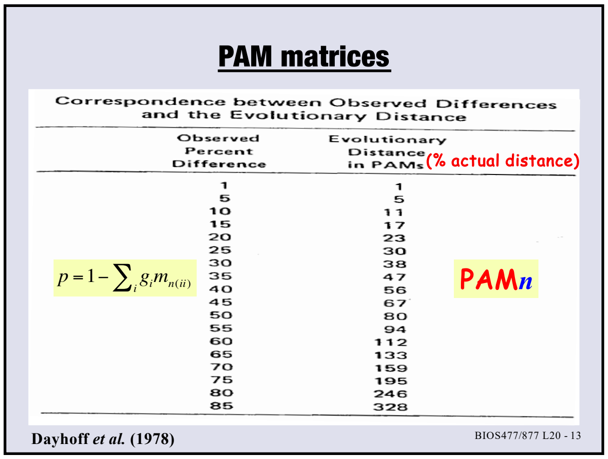
[Matrix components]
$m_{n(ij)}$: Probability of AA$_j$ replaced by AA$_i$
$m_{n(ii)}$: Probability of AA$_i$ not changing

M1		A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly
A	Ala	9867	2	9	10	3	8	17	21
R	Arg	1	9913	1	0	1	10	0	0
N	Asn	4	1	9822	36	0	4	6	6
D	Asp	6	0	42	9859	0	6	53	6
C	Cys	1	1	0	0	9973	0	0	0
Q	Gln	3	9	4	5	0	9876	27	1
E	Glu	10	0	7	56	0	35	9865	4
G	Gly	21	1	12	11	1	3	7	9935

$$p=1-\sum_i g_i m_{n(ii)}$$

p: the proportion of amino acid differences when two sequences have n% distance
g_i: the equilibrium frequency of the amino acid i

12



PAM matrices

Correspondence between Observed Differences and the Evolutionary Distance

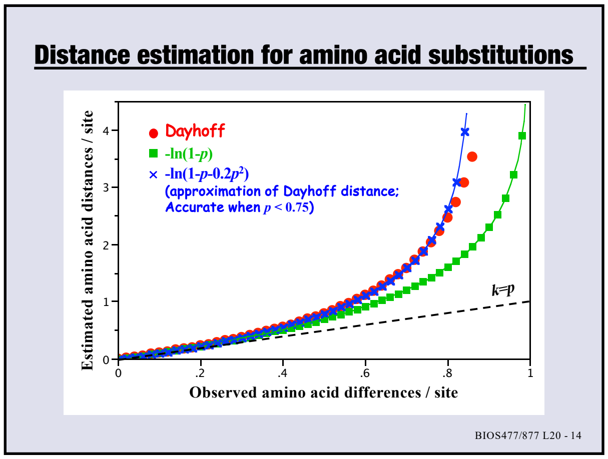
Observed Percent Difference	Evolutionary Distance in PAMs (% actual distance)
1	1
5	5
10	11
15	17
20	23
25	30
30	38
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328

$p = 1 - \sum_i g_i m_{n(ii)}$

PAM*n*

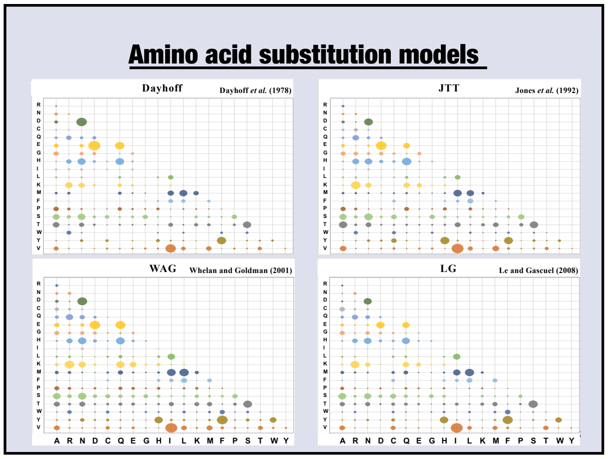
Dayhoff *et al.* (1978)

13



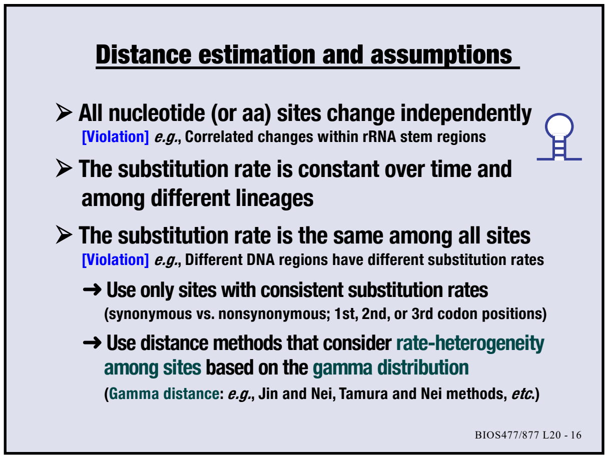
Distance estimation for amino acid substitutions

14



Amino acid substitution models

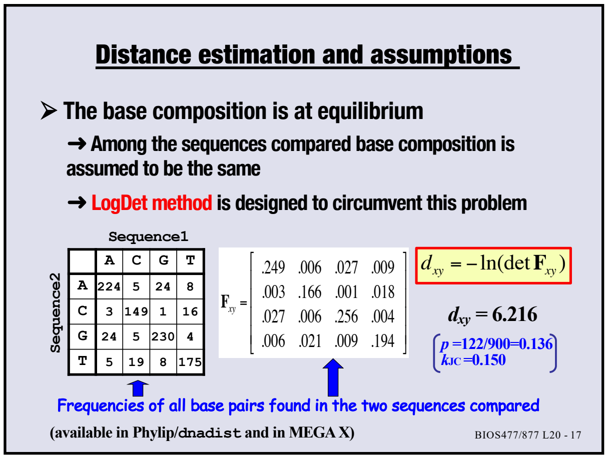
15



Distance estimation and assumptions

- All nucleotide (or aa) sites change independently
  [Violation] *e.g.*, Correlated changes within rRNA stem regions
- The substitution rate is constant over time and among different lineages
- The substitution rate is the same among all sites
  [Violation] *e.g.*, Different DNA regions have different substitution rates
  - → Use only sites with consistent substitution rates
    (synonymous vs. nonsynonymous; 1st, 2nd, or 3rd codon positions)
  - → Use distance methods that consider rate-heterogeneity among sites based on the gamma distribution
    (Gamma distance: *e.g.*, Jin and Nei, Tamura and Nei methods, *etc.*)

16



Distance estimation and assumptions

- The base composition is at equilibrium
  - → Among the sequences compared base composition is assumed to be the same
  - → LogDet method is designed to circumvent this problem

Sequence1

Sequence2	A	C	G	T
A	224	5	24	8
C	3	149	1	16
G	24	5	230	4
T	5	19	8	175

$$\mathbf{F}_{xy} = \begin{bmatrix} .249 & .006 & .027 & .009 \\ .003 & .166 & .001 & .018 \\ .027 & .006 & .256 & .004 \\ .006 & .021 & .009 & .194 \end{bmatrix}$$

$$\boxed{d_{xy} = -\ln(\det \mathbf{F}_{xy})}$$

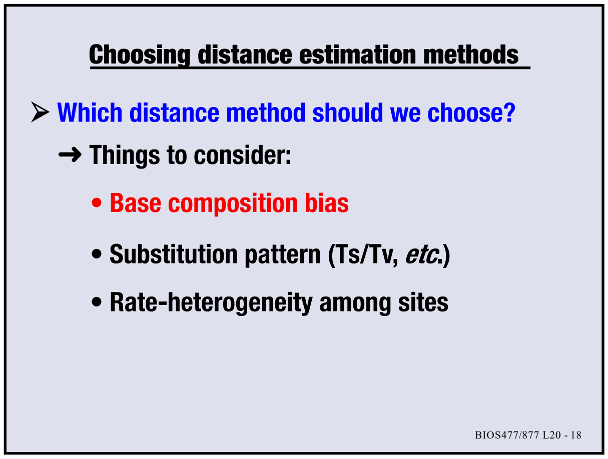
$$d_{xy} = 6.216$$

$$\begin{pmatrix} p = 122/900 = 0.136 \\ k_{JC} = 0.150 \end{pmatrix}$$

Frequencies of all base pairs found in the two sequences compared

(available in Phylip/`dnadist` and in MEGA X)

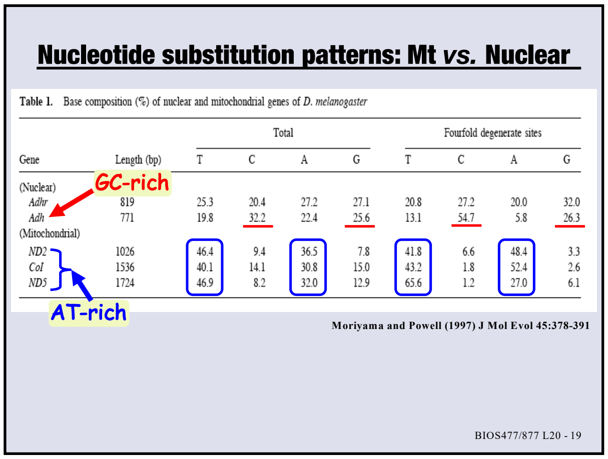
17



Choosing distance estimation methods

- Which distance method should we choose?
  - → Things to consider:
    - Base composition bias
    - Substitution pattern (Ts/Tv, *etc.*)
    - Rate-heterogeneity among sites

18



Nucleotide substitution patterns: Mt *vs.* Nuclear

Table 1. Base composition (%) of nuclear and mitochondrial genes of *D. melanogaster*

Gene	Length (bp)	Total				Fourfold degenerate sites			
		T	C	A	G	T	C	A	G
(Nuclear)									
Adhr	819	25.3	20.4	27.2	27.1	20.8	27.2	20.0	32.0
Adh	771	19.8	32.2	22.4	25.6	13.1	54.7	5.8	26.3
(Mitochondrial)									
ND2	1026	46.4	9.4	36.5	7.8	41.8	6.6	48.4	3.3
CoI	1536	40.1	14.1	30.8	15.0	43.2	1.8	52.4	2.6
ND5	1724	46.9	8.2	32.0	12.9	65.6	1.2	27.0	6.1

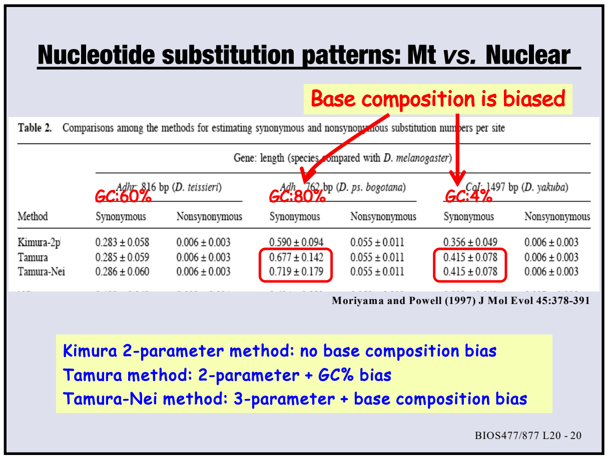
GC-rich

AT-rich

Moriyama and Powell (1997) J Mol Evol 45:378-391

BIOS477/877 L20 - 19

19



Nucleotide substitution patterns: Mt *vs.* Nuclear

Base composition is biased

Table 2. Comparisons among the methods for estimating synonymous and nonsynonymous substitution numbers per site

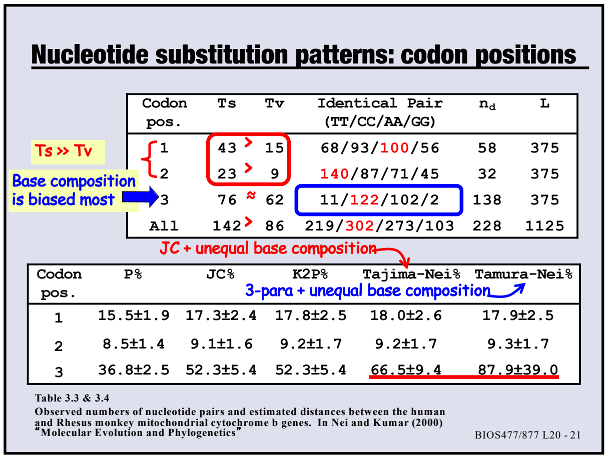
Method	Gene: length (species compared with *D. melanogaster*)					
	Adhr: 816 bp (*D. teissieri*) GC:60%		*Adh*: 762 bp (*D. ps. bogotana*) GC:80%		*CoI*: 1497 bp (*D. yakuba*) GC:4%	
	Synonymous	Nonsynonymous	Synonymous	Nonsynonymous	Synonymous	Nonsynonymous
Kimura-2p	0.283 ± 0.058	0.006 ± 0.003	0.590 ± 0.094	0.055 ± 0.011	0.356 ± 0.049	0.006 ± 0.003
Tamura	0.285 ± 0.059	0.006 ± 0.003	0.677 ± 0.142	0.055 ± 0.011	0.415 ± 0.078	0.006 ± 0.003
Tamura-Nei	0.286 ± 0.060	0.006 ± 0.003	0.719 ± 0.179	0.055 ± 0.011	0.415 ± 0.078	0.006 ± 0.003

Moriyama and Powell (1997) J Mol Evol 45:378-391

Kimura 2-parameter method: no base composition bias
Tamura method: 2-parameter + GC% bias
Tamura-Nei method: 3-parameter + base composition bias

BIOS477/877 L20 - 20

20



Nucleotide substitution patterns: codon positions

Ts >> Tv

Base composition is biased most

Codon pos.	Ts		Tv	Identical Pair (TT/CC/AA/GG)	n_d	L
1	43	>	15	68/93/100/56	58	375
2	23	>	9	140/87/71/45	32	375
3	76	≈	62	11/122/102/2	138	375
All	142	>	86	219/302/273/103	228	1125

JC + unequal base composition

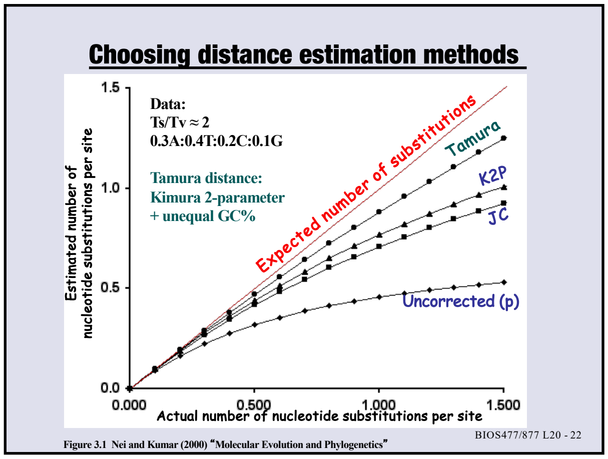
3-para + unequal base composition

Codon pos.	P%	JC%	K2P%	Tajima-Nei%	Tamura-Nei%
1	15.5±1.9	17.3±2.4	17.8±2.5	18.0±2.6	17.9±2.5
2	8.5±1.4	9.1±1.6	9.2±1.7	9.2±1.7	9.3±1.7
3	36.8±2.5	52.3±5.4	52.3±5.4	66.5±9.4	87.9±39.0

Table 3.3 & 3.4
Observed numbers of nucleotide pairs and estimated distances between the human and Rhesus monkey mitochondrial cytochrome b genes. In Nei and Kumar (2000) "Molecular Evolution and Phylogenetics"

BIOS477/877 L20 - 21

21

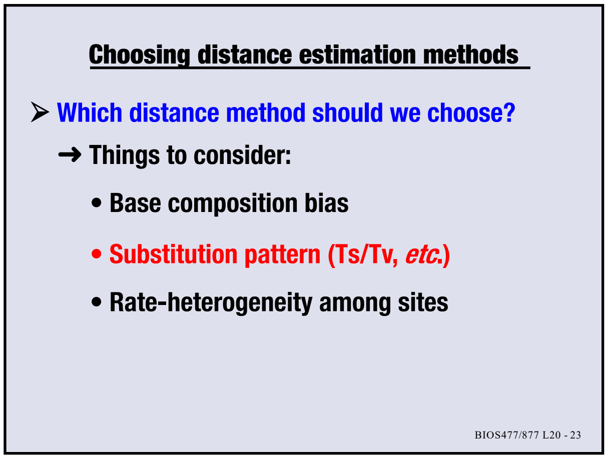


Choosing distance estimation methods

Figure 3.1 Nei and Kumar (2000) "Molecular Evolution and Phylogenetics"

BIOS477/877 L20 - 22

22

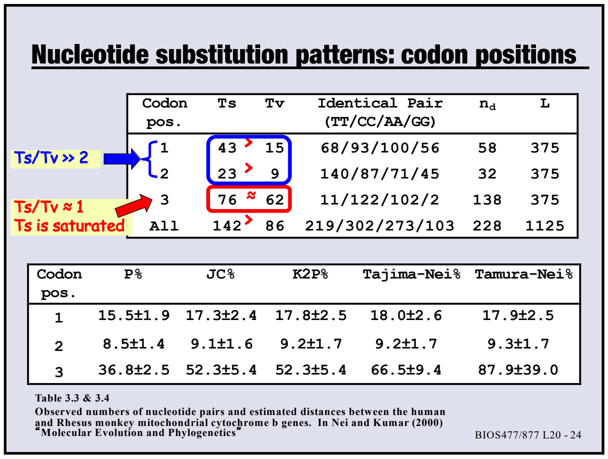


Choosing distance estimation methods

- Which distance method should we choose?
  - ➜ Things to consider:
    - Base composition bias
    - Substitution pattern (Ts/Tv, *etc.*)
    - Rate-heterogeneity among sites

BIOS477/877 L20 - 23

23



Nucleotide substitution patterns: codon positions

Ts/Tv >> 2

Ts/Tv ≈ 1
Ts is saturated

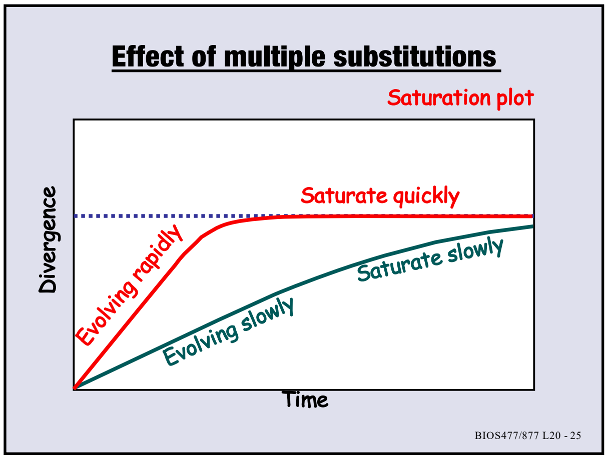
Codon pos.	Ts		Tv	Identical Pair (TT/CC/AA/GG)	n_d	L
1	43	>	15	68/93/100/56	58	375
2	23	>	9	140/87/71/45	32	375
3	76	≈	62	11/122/102/2	138	375
All	142	>	86	219/302/273/103	228	1125

Codon pos.	P%	JC%	K2P%	Tajima-Nei%	Tamura-Nei%
1	15.5±1.9	17.3±2.4	17.8±2.5	18.0±2.6	17.9±2.5
2	8.5±1.4	9.1±1.6	9.2±1.7	9.2±1.7	9.3±1.7
3	36.8±2.5	52.3±5.4	52.3±5.4	66.5±9.4	87.9±39.0

Table 3.3 & 3.4
Observed numbers of nucleotide pairs and estimated distances between the human and Rhesus monkey mitochondrial cytochrome b genes. In Nei and Kumar (2000) "Molecular Evolution and Phylogenetics"

BIOS477/877 L20 - 24

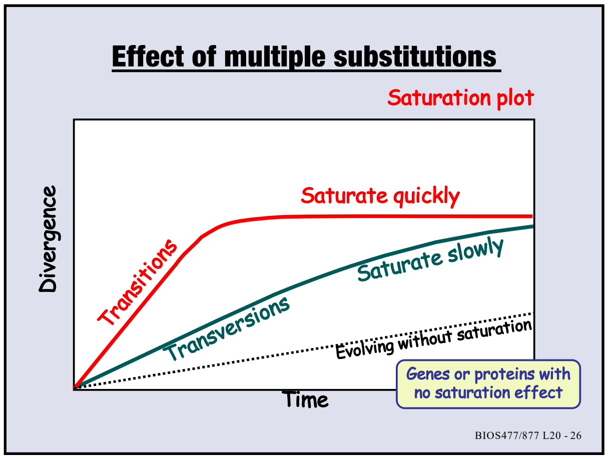
24



Effect of multiple substitutions

Saturation plot

25

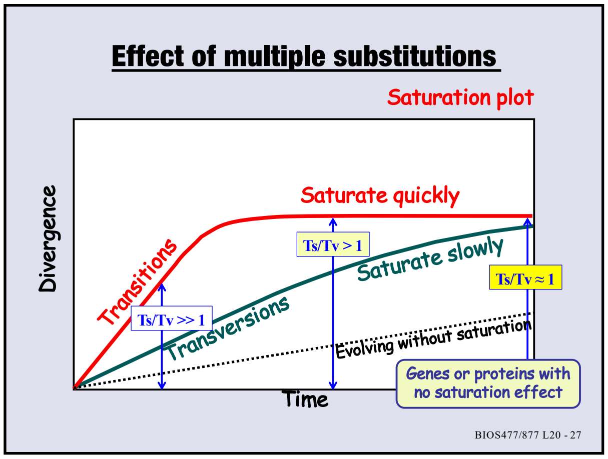


Effect of multiple substitutions

Saturation plot

Genes or proteins with no saturation effect

26

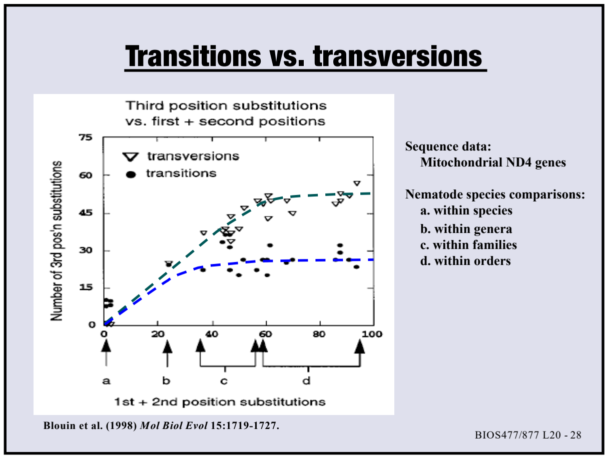


Effect of multiple substitutions

Saturation plot

Genes or proteins with no saturation effect

27



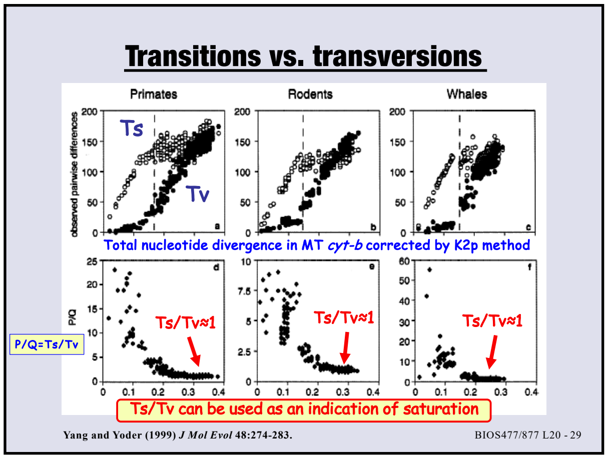
Transitions vs. transversions

Sequence data:
Mitochondrial ND4 genes

Nematode species comparisons:
a. within species
b. within genera
c. within families
d. within orders

Blouin et al. (1998) *Mol Biol Evol* 15:1719-1727.

28



Transitions vs. transversions

Total nucleotide divergence in MT *cyt-b* corrected by K2p method

P/Q=Ts/Tv

Ts/Tv can be used as an indication of saturation

Yang and Yoder (1999) *J Mol Evol* 48:274-283.

29

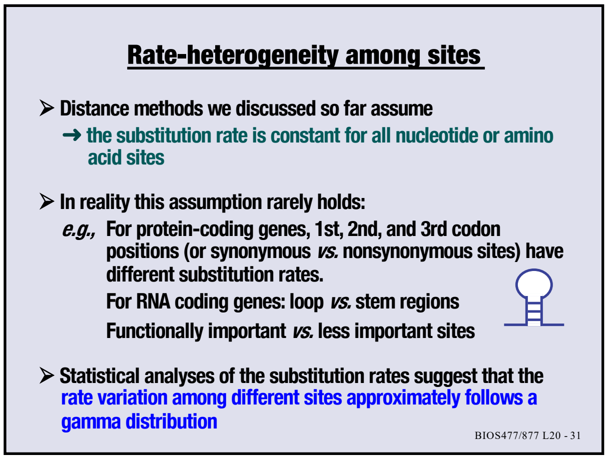
Choosing distance estimation methods

➢ **Which distance method should we choose?**

➜ **Things to consider:**

- **Base composition bias**
- **Substitution pattern (Ts/Tv, *etc.*)**
- **Rate-heterogeneity among sites**

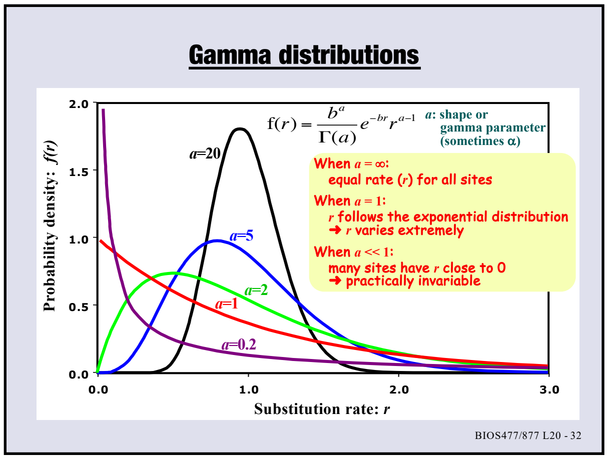
30



Rate-heterogeneity among sites

- Distance methods we discussed so far assume
  ➜ the substitution rate is constant for all nucleotide or amino acid sites
- In reality this assumption rarely holds:
  e.g., For protein-coding genes, 1st, 2nd, and 3rd codon positions (or synonymous *vs.* nonsynonymous sites) have different substitution rates.
  For RNA coding genes: loop *vs.* stem regions
  Functionally important *vs.* less important sites
- Statistical analyses of the substitution rates suggest that the rate variation among different sites approximately follows a gamma distribution

31



Gamma distributions

$$f(r) = \frac{b^a}{\Gamma(a)} e^{-br} r^{a-1}$$

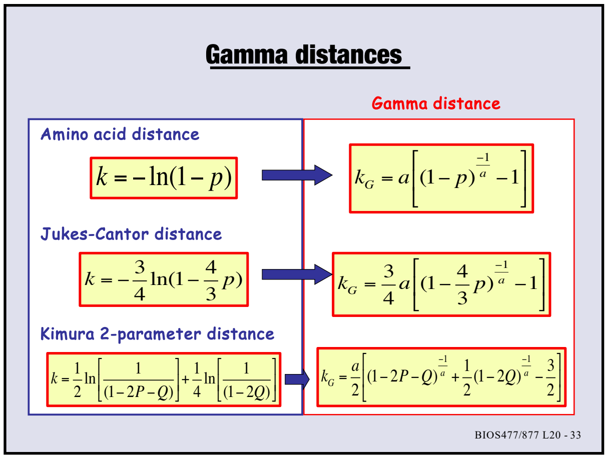
a: shape or gamma parameter (sometimes α)

When $a = \infty$:
equal rate (r) for all sites

When $a = 1$:
r follows the exponential distribution
➜ r varies extremely

When $a << 1$:
many sites have r close to 0
➜ practically invariable

32



Gamma distances

Gamma distance

Amino acid distance

$$\boxed{k = -\ln(1-p)} \Rightarrow \boxed{k_G = a\left[(1-p)^{\frac{-1}{a}} - 1\right]}$$

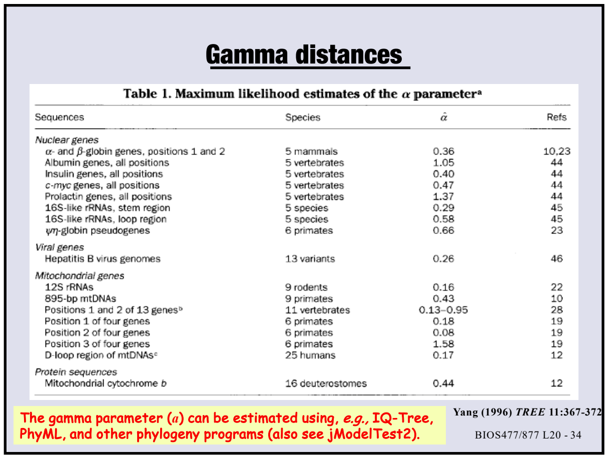
Jukes-Cantor distance

$$\boxed{k = -\frac{3}{4}\ln(1-\frac{4}{3}p)} \Rightarrow \boxed{k_G = \frac{3}{4}a\left[(1-\frac{4}{3}p)^{\frac{-1}{a}} - 1\right]}$$

Kimura 2-parameter distance

$$\boxed{k = \frac{1}{2}\ln\left[\frac{1}{(1-2P-Q)}\right] + \frac{1}{4}\ln\left[\frac{1}{(1-2Q)}\right]} \Rightarrow \boxed{k_G = \frac{a}{2}\left[(1-2P-Q)^{\frac{-1}{a}} + \frac{1}{2}(1-2Q)^{\frac{-1}{a}} - \frac{3}{2}\right]}$$

33



Gamma distances

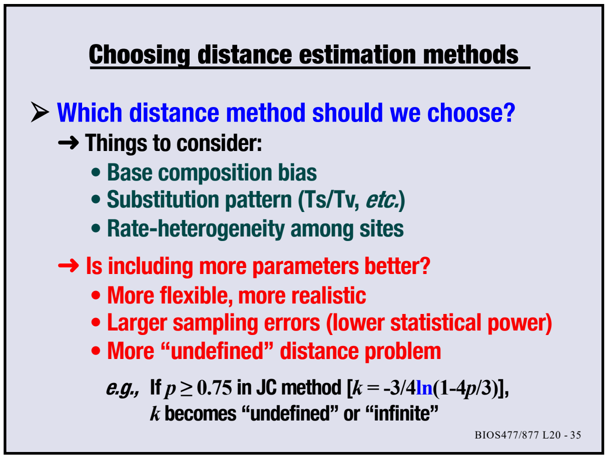
Table 1. Maximum likelihood estimates of the α parameter[a]

Sequences	Species	$\hat{\alpha}$	Refs
Nuclear genes			
α- and β-globin genes, positions 1 and 2	5 mammals	0.36	10,23
Albumin genes, all positions	5 vertebrates	1.05	44
Insulin genes, all positions	5 vertebrates	0.40	44
c-myc genes, all positions	5 vertebrates	0.47	44
Prolactin genes, all positions	5 vertebrates	1.37	44
16S-like rRNAs, stem region	5 species	0.29	45
16S-like rRNAs, loop region	5 species	0.58	45
$\psi\eta$-globin pseudogenes	6 primates	0.66	23
Viral genes			
Hepatitis B virus genomes	13 variants	0.26	46
Mitochondrial genes			
12S rRNAs	9 rodents	0.16	22
895-bp mtDNAs	9 primates	0.43	10
Positions 1 and 2 of 13 genes[b]	11 vertebrates	0.13–0.95	28
Position 1 of four genes	6 primates	0.18	19
Position 2 of four genes	6 primates	0.08	19
Position 3 of four genes	6 primates	1.58	19
D-loop region of mtDNAs[c]	25 humans	0.17	12
Protein sequences			
Mitochondrial cytochrome *b*	16 deuterostomes	0.44	12

The gamma parameter (a) can be estimated using, *e.g.,* IQ-Tree, PhyML, and other phylogeny programs (also see jModelTest2).

Yang (1996) ***TREE*** **11:367-372**

34



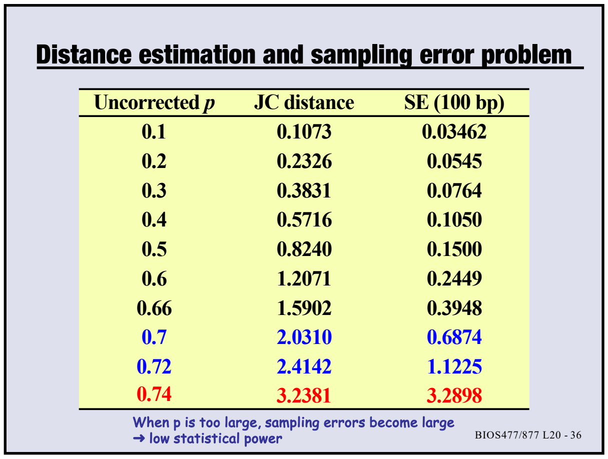
Choosing distance estimation methods

- Which distance method should we choose?
  ➜ Things to consider:
  - Base composition bias
  - Substitution pattern (Ts/Tv, *etc.*)
  - Rate-heterogeneity among sites

  ➜ Is including more parameters better?
  - More flexible, more realistic
  - Larger sampling errors (lower statistical power)
  - More “undefined” distance problem

  e.g., If $p \geq 0.75$ in JC method [$k = -3/4\ln(1-4p/3)$],
  k becomes “undefined” or “infinite”

35

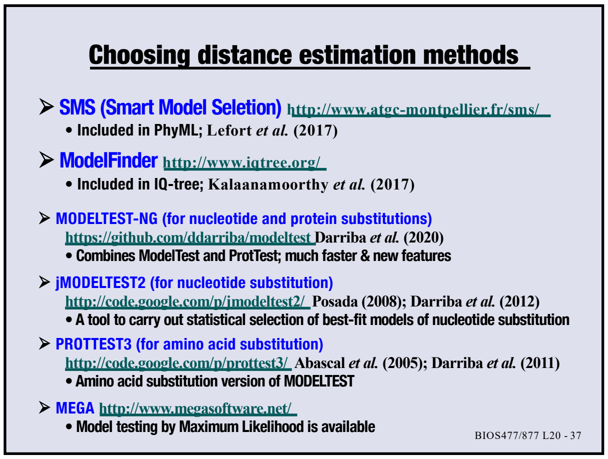


Distance estimation and sampling error problem

Uncorrected *p*	JC distance	SE (100 bp)
0.1	0.1073	0.03462
0.2	0.2326	0.0545
0.3	0.3831	0.0764
0.4	0.5716	0.1050
0.5	0.8240	0.1500
0.6	1.2071	0.2449
0.66	1.5902	0.3948
0.7	2.0310	0.6874
0.72	2.4142	1.1225
0.74	3.2381	3.2898

When p is too large, sampling errors become large
➜ low statistical power

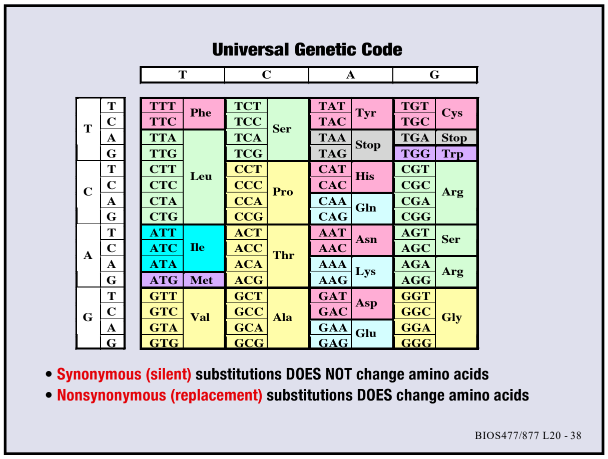
36



Choosing distance estimation methods

- **SMS (Smart Model Seletion)** http://www.atgc-montpellier.fr/sms/
  - Included in PhyML; Lefort *et al.* (2017)
- **ModelFinder** http://www.iqtree.org/
  - Included in IQ-tree; Kalaanamoorthy *et al.* (2017)
- **MODELTEST-NG (for nucleotide and protein substitutions)**
  https://github.com/ddarriba/modeltest Darriba *et al.* (2020)
  - Combines ModelTest and ProtTest; much faster & new features
- **jMODELTEST2 (for nucleotide substitution)**
  http://code.google.com/p/jmodeltest2/ Posada (2008); Darriba *et al.* (2012)
  - A tool to carry out statistical selection of best-fit models of nucleotide substitution
- **PROTTEST3 (for amino acid substitution)**
  http://code.google.com/p/prottest3/ Abascal *et al.* (2005); Darriba *et al.* (2011)
  - Amino acid substitution version of MODELTEST
- **MEGA** http://www.megasoftware.net/
  - Model testing by Maximum Likelihood is available

37

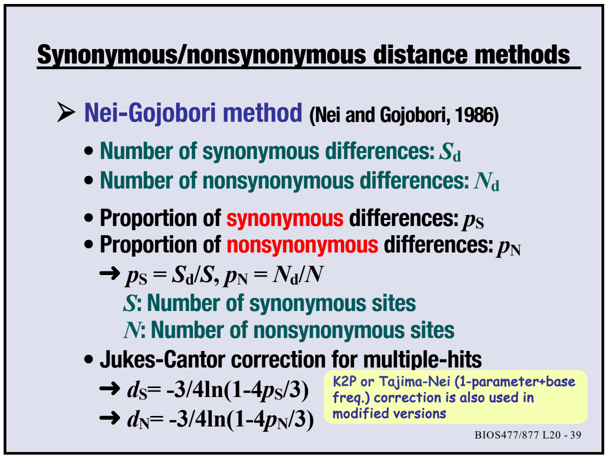


Universal Genetic Code

		T		C		A		G	
T	T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
	C	TTC		TCC		TAC		TGC	
	A	TTA	Leu	TCA		TAA	Stop	TGA	Stop
	G	TTG		TCG		TAG		TGG	Trp
C	T	CTT		CCT	Pro	CAT	His	CGT	Arg
	C	CTC		CCC		CAC		CGC	
	A	CTA		CCA		CAA	Gln	CGA	
	G	CTG		CCG		CAG		CGG	
A	T	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
	C	ATC		ACC		AAC		AGC	
	A	ATA		ACA		AAA	Lys	AGA	Arg
	G	ATG	Met	ACG		AAG		AGG	
G	T	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
	C	GTC		GCC		GAC		GGC	
	A	GTA		GCA		GAA	Glu	GGA	
	G	GTG		GCG		GAG		GGG	

- Synonymous (silent) substitutions DOES NOT change amino acids
- Nonsynonymous (replacement) substitutions DOES change amino acids

38



Synonymous/nonsynonymous distance methods

- **Nei-Gojobori method** (Nei and Gojobori, 1986)
  - Number of synonymous differences: S_d
  - Number of nonsynonymous differences: N_d
  - Proportion of synonymous differences: p_S
  - Proportion of nonsynonymous differences: p_N

    ➜ $p_S = S_d/S$, $p_N = N_d/N$

    S: Number of synonymous sites

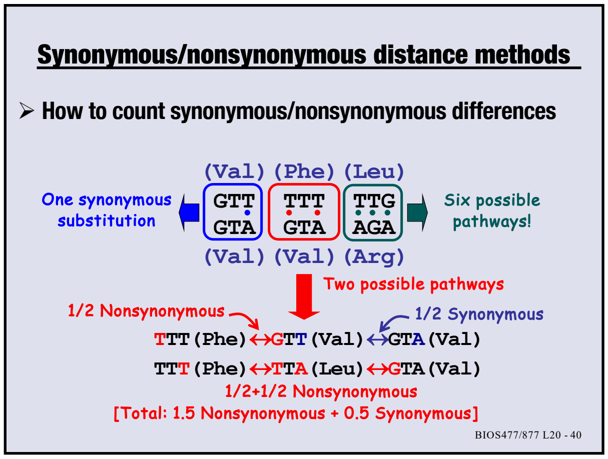
    N: Number of nonsynonymous sites
  - Jukes-Cantor correction for multiple-hits

    ➜ $d_S = -3/4 \ln(1-4p_S/3)$

    ➜ $d_N = -3/4 \ln(1-4p_N/3)$

K2P or Tajima-Nei (1-parameter+base freq.) correction is also used in modified versions

39



Synonymous/nonsynonymous distance methods

- How to count synonymous/nonsynonymous differences

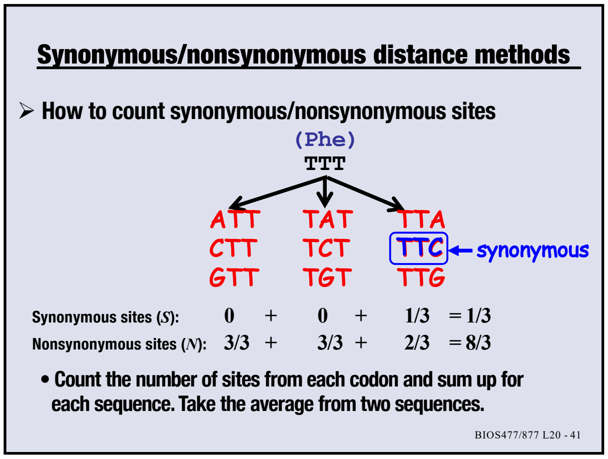
TTT (Phe) ↔ GTT (Val) ↔ GTA (Val)

TTT (Phe) ↔ TTA (Leu) ↔ GTA (Val)

1/2+1/2 Nonsynonymous

[Total: 1.5 Nonsynonymous + 0.5 Synonymous]

40



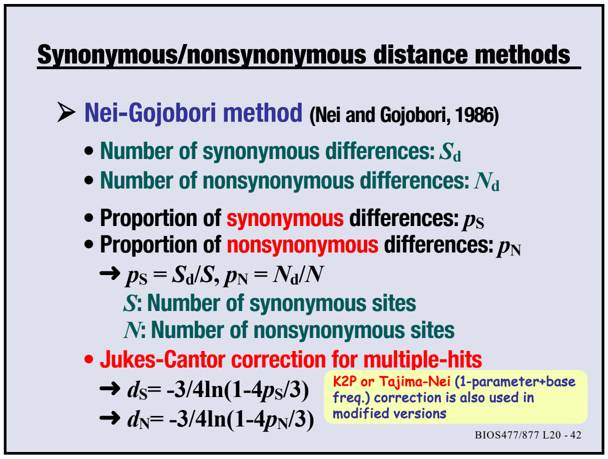
Synonymous/nonsynonymous distance methods

- How to count synonymous/nonsynonymous sites

Synonymous sites (S):	0	+	0	+	1/3	= 1/3
Nonsynonymous sites (N):	3/3	+	3/3	+	2/3	= 8/3

- Count the number of sites from each codon and sum up for each sequence. Take the average from two sequences.

41



Synonymous/nonsynonymous distance methods

- **Nei-Gojobori method** (Nei and Gojobori, 1986)
  - Number of synonymous differences: S_d
  - Number of nonsynonymous differences: N_d
  - Proportion of synonymous differences: p_S
  - Proportion of nonsynonymous differences: p_N

    ➜ $p_S = S_d/S$, $p_N = N_d/N$

    S: Number of synonymous sites

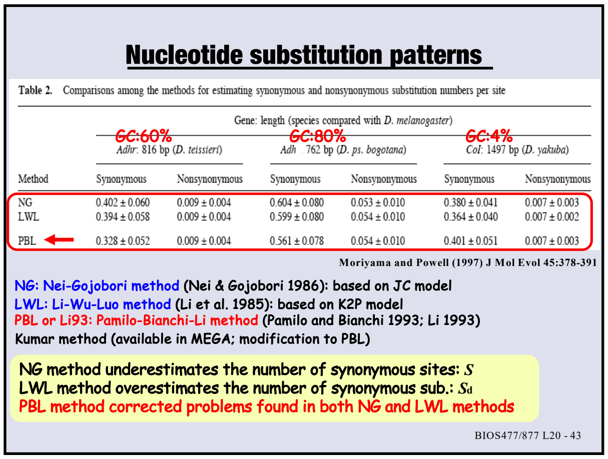
    N: Number of nonsynonymous sites
  - Jukes-Cantor correction for multiple-hits

    ➜ $d_S = -3/4 \ln(1-4p_S/3)$

    ➜ $d_N = -3/4 \ln(1-4p_N/3)$

K2P or Tajima-Nei (1-parameter+base freq.) correction is also used in modified versions

42



Nucleotide substitution patterns

Table 2. Comparisons among the methods for estimating synonymous and nonsynonymous substitution numbers per site

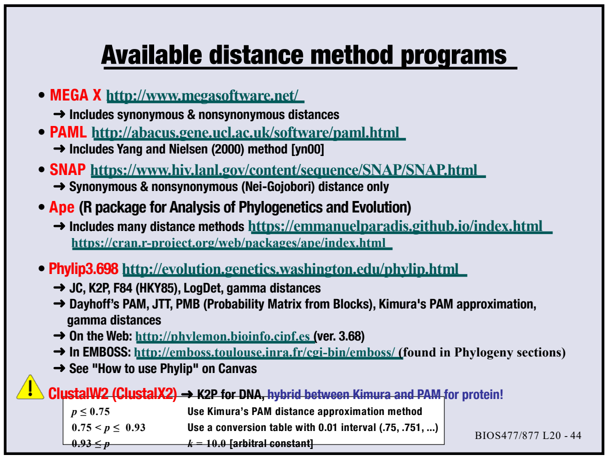
	Gene: length (species compared with *D. melanogaster*)					
	GC:60% *Adhr*: 816 bp (*D. teissieri*)		GC:80% *Adh* 762 bp (*D. ps. bogotana*)		GC:4% *CoI*: 1497 bp (*D. yakuba*)	
Method	Synonymous	Nonsynonymous	Synonymous	Nonsynonymous	Synonymous	Nonsynonymous
NG	0.402 ± 0.060	0.009 ± 0.004	0.604 ± 0.080	0.053 ± 0.010	0.380 ± 0.041	0.007 ± 0.003
LWL	0.394 ± 0.058	0.009 ± 0.004	0.599 ± 0.080	0.054 ± 0.010	0.364 ± 0.040	0.007 ± 0.002
PBL ←	0.328 ± 0.052	0.009 ± 0.004	0.561 ± 0.078	0.054 ± 0.010	0.401 ± 0.051	0.007 ± 0.003

Moriyama and Powell (1997) J Mol Evol 45:378-391

NG: Nei-Gojobori method (Nei & Gojobori 1986): based on JC model
LWL: Li-Wu-Luo method (Li et al. 1985): based on K2P model
PBL or Li93: Pamilo-Bianchi-Li method (Pamilo and Bianchi 1993; Li 1993)
Kumar method (available in MEGA; modification to PBL)

NG method underestimates the number of synonymous sites: S
LWL method overestimates the number of synonymous sub.: S_d
PBL method corrected problems found in both NG and LWL methods

43



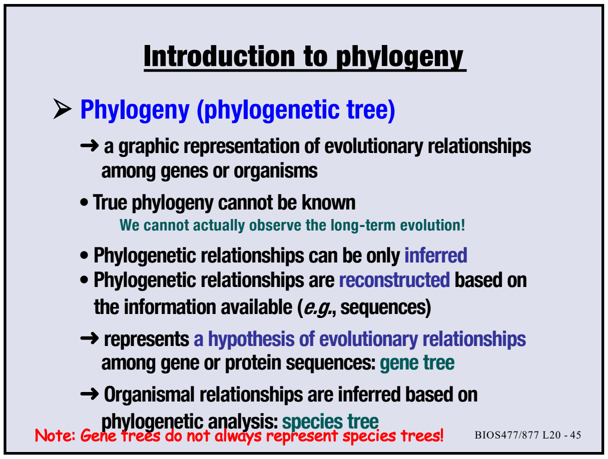
Available distance method programs

- MEGA X http://www.megasoftware.net/
  - ➜ Includes synonymous & nonsynonymous distances
- PAML http://abacus.gene.ucl.ac.uk/software/paml.html
  - ➜ Includes Yang and Nielsen (2000) method [yn00]
- SNAP https://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html
  - ➜ Synonymous & nonsynonymous (Nei-Gojobori) distance only
- Ape (R package for Analysis of Phylogenetics and Evolution)
  - ➜ Includes many distance methods https://emmanuelparadis.github.io/index.html https://cran.r-project.org/web/packages/ape/index.html
- Phylip3.698 http://evolution.genetics.washington.edu/phylip.html
  - ➜ JC, K2P, F84 (HKY85), LogDet, gamma distances
  - ➜ Dayhoff's PAM, JTT, PMB (Probability Matrix from Blocks), Kimura's PAM approximation, gamma distances
  - ➜ On the Web: http://phylemon.bioinfo.cipf.es (ver. 3.68)
  - ➜ In EMBOSS: http://emboss.toulouse.inra.fr/cgi-bin/emboss/ (found in Phylogeny sections)
  - ➜ See "How to use Phylip" on Canvas

⚠ ClustalW2 (ClustalX2) ➜ K2P for DNA, hybrid between Kimura and PAM for protein!

$p \le 0.75$	Use Kimura's PAM distance approximation method
$0.75 < p \le 0.93$	Use a conversion table with 0.01 interval (.75, .751, ...)
$0.93 \le p$	$k = 10.0$ [arbitral constant]

44



Introduction to phylogeny

➢ Phylogeny (phylogenetic tree)

➜ a graphic representation of evolutionary relationships among genes or organisms

- True phylogeny cannot be known

  We cannot actually observe the long-term evolution!

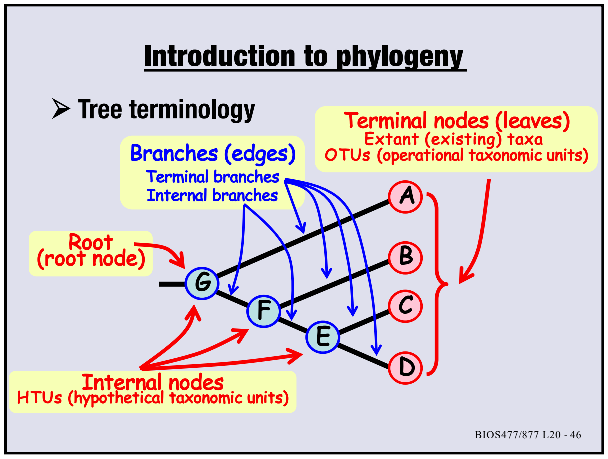
- Phylogenetic relationships can be only inferred
- Phylogenetic relationships are reconstructed based on the information available (*e.g.*, sequences)

➜ represents a hypothesis of evolutionary relationships among gene or protein sequences: gene tree

➜ Organismal relationships are inferred based on phylogenetic analysis: species tree

Note: Gene trees do not always represent species trees!

45



Introduction to phylogeny

➢ Tree terminology

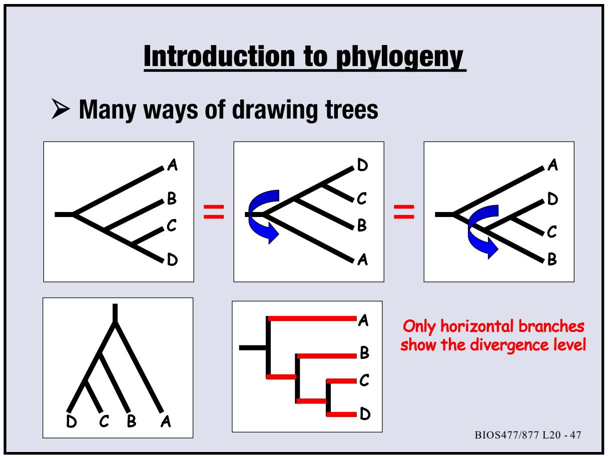
Terminal nodes (leaves)
Extant (existing) taxa
OTUs (operational taxonomic units)

Branches (edges)
Terminal branches
Internal branches

Root
(root node)

Internal nodes
HTUs (hypothetical taxonomic units)

46

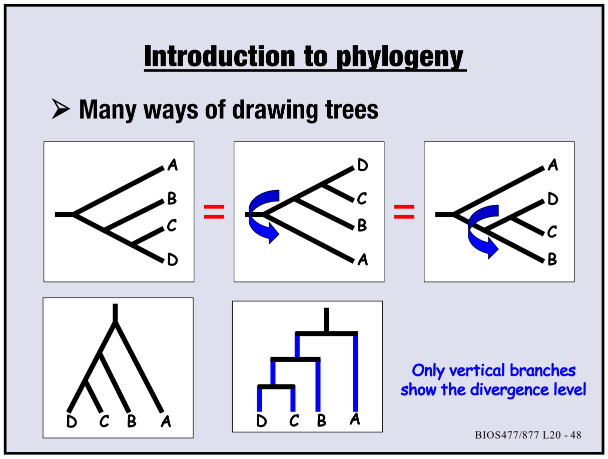


Introduction to phylogeny

➢ Many ways of drawing trees

Only horizontal branches show the divergence level

47

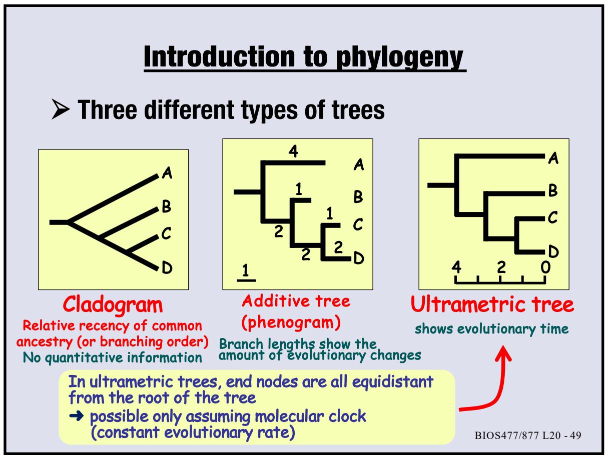


Introduction to phylogeny

➢ Many ways of drawing trees

Only vertical branches show the divergence level

48



Introduction to phylogeny

➢ Three different types of trees

Cladogram

Relative recency of common ancestry (or branching order)

No quantitative information

Additive tree (phenogram)

Branch lengths show the amount of evolutionary changes

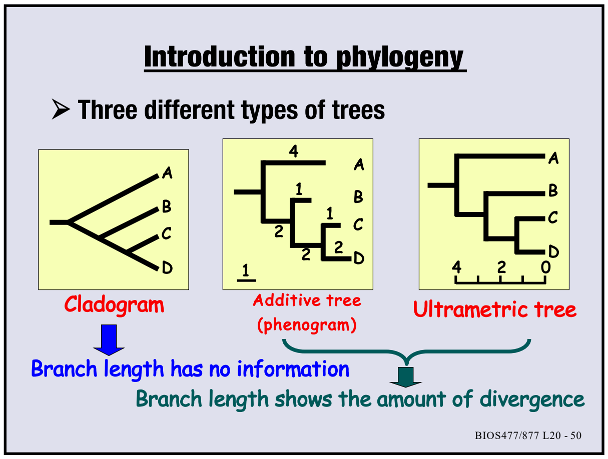
Ultrametric tree

shows evolutionary time

In ultrametric trees, end nodes are all equidistant from the root of the tree
➜ possible only assuming molecular clock (constant evolutionary rate)

BIOS477/877 L20 - 49

49



Introduction to phylogeny

➢ Three different types of trees

Cladogram

Additive tree (phenogram)

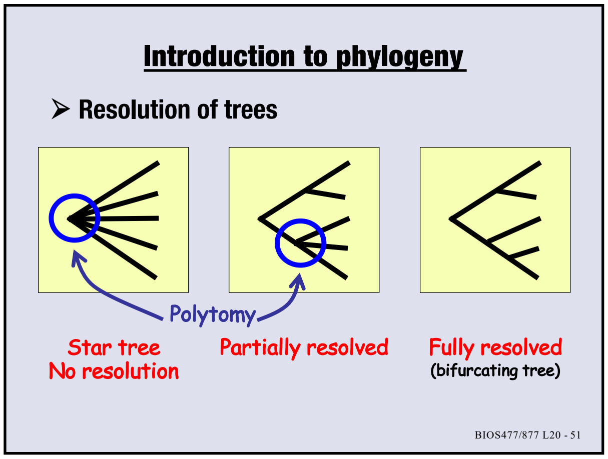
Ultrametric tree

Branch length has no information

Branch length shows the amount of divergence

BIOS477/877 L20 - 50

50



Introduction to phylogeny

➢ Resolution of trees

Polytomy

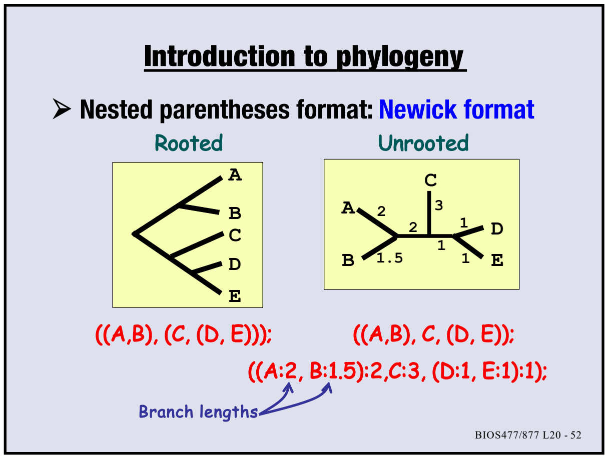
Star tree
No resolution

Partially resolved

Fully resolved
(bifurcating tree)

BIOS477/877 L20 - 51

51



Introduction to phylogeny

➢ Nested parentheses format: Newick format

Rooted

Unrooted

((A,B), (C, (D, E)));

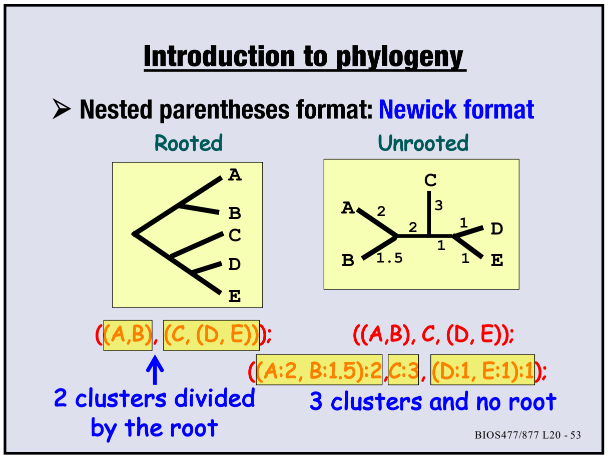
((A,B), C, (D, E));

((A:2, B:1.5):2,C:3, (D:1, E:1):1);

Branch lengths

BIOS477/877 L20 - 52

52



Introduction to phylogeny

➢ Nested parentheses format: Newick format

Rooted

Unrooted

((A,B), (C, (D, E)));

((A,B), C, (D, E));

((A:2, B:1.5):2,C:3, (D:1, E:1):1);

2 clusters divided by the root

3 clusters and no root

BIOS477/877 L20 - 53

53