Serik Sagitov, Chalmers Tekniska Högskola, November 2, 2003

4. Molecular population genetics

Polymorphisms on the amino acid and nucleotide level

- 4.1 theories of molecular evolution
- 4.2 sequence divergence rates
- 4.3 rates corrected for multiple hits
- 4.4 molecular clocks
- 4.5 gene genealogy and coalescent
- 4.6 within species mol. polymorphism
- 4.7 synonimous and non-synonimous rates
- 4.8 two tests of neutrality
- 4.9 recombination and polymorhism

4.1 Theories of molecular evolution

most mutations are deleterious and quickly removed

Classical theory

- natural selection is the major evolutionary force
- predicts little genetic variation
- because positive mutations are quickly fixed

Balance theory

- most polymorphisms due to balanced selection
- fails to explain protein electrophoresis results
- 15-50% of enzyme coding genes are polymorphic
- with two or more widespread alleles

Neutral theory

most polymorphisms are nearly selectively neutral

RGD is a major evolutionary force

Ex 1: heterozygosity and population size

neutral theory prediction for IAM: $\hat{H} = \frac{\theta}{1+\theta}, \theta = 4N_e\mu$

μ = mutation rate per nucleotide site per generation

Fig 8.2, p. 319: 77 species data do not fit the prediction

variation in H is lower than expected under neutrality given the huge variation in EPS

Possible explanations

several evolutionary forces involved

different species - different magnitudes of the forces

incorrectly estimated N_e

4.2 Sequence divergence rates

Two homologous sequences

sequence length: L amino acids, $l = 3L$ nucleotide sites

d = observed nucleotide differences per site, $0 \leq d \leq 1$

D = observed amino acid diff. per site, $0 \leq D \leq 1$

t = divergence time between the homologuos sequences

Parameter estimation problem: using d, D estimate

nucleotide substitution rate $\lambda = \frac{k}{2t}$

amino acid replacement rate $\Lambda = \frac{K}{2t}$

k, K = actual numbers of differences per site

Multiple hits examples

1) observed A $\to$ C, full history A $\to$ T $\to$ G $\to$ C

2) observed A $\to$ A, full history A $\to$ T $\to$ A

Ex 2: bacterial gene

Coding region of *trpA* in two related bacterial strains

K12 (*E.coli*) and LT2 (*Salmonella typhimurium*)

diverged $t = 80$ MY ago (mammalian radiation)

0*04	004*	004*	002	002	002	002	004	004	002
GTC	GCA	CCT	ATC	TTC	ATC	TGC	CCG	CCA	AAT
Val	Ala	Pro	Ile	Phe	Ile	Cys	Pro	Pro	Asn
ATC	GCG	CCG	ATC	TTC	ATC	TGC	CCG	CCA	AAT
Ile	Ala	Pro	Ile	Phe	Ile	Cys	Pro	Pro	Asn
N	S	S							

004*	002	002	002*	204*	204	004	002	0*02*	004*
GCC	GAT	GAC	GAC	CTG	CTG	CGC	CAG	ATA	GCC
Ala	Asp	Asp	Asp	Leu	Leu	Arg	Gln	Ile	Ala
GCG	GAT	GAC	GAT	CTT	CTG	CGC	CAG	GTC	GCA
Ala	Asp	Asp	Asp	Leu	Leu	Arg	Gln	Val	Ala
S			S	S				N S	S

Observed differences: 9 nucleotide, 2 amino acid

$l = 60$, $d = 9/60 = 0.15$, $L = 20$, $D = 2/20 = 0.10$

Uncorrected estimates of the rates λ and Λ:

$\tilde{\lambda} = \frac{d}{2t} = 0.94 \cdot 10^{-9}$ substitutions per site per year

$\tilde{\Lambda} = \frac{D}{2t} = 0.63 \cdot 10^{-9}$ replacements per site per year

Substitution and mutation rates

Diffusion approximation prediction of λ

$\lambda = \#(\text{mutations per gener}) \times (\text{fixation probability})$

$= 2N\mu \times u(\frac{1}{2N}) \approx \frac{4N_e s\mu}{1-e^{-4N_e s}}$ (additive selection)

If most substitutions are

deleterious: λ decreases with N_e

advantageous: λ increases with N_e

Neutral substitutions: $\lambda = \mu$ is independent of N_e

Ex 3: diffusion simulations

Fig 8.1, p. 317: neutral substitutions for different μ

average fixation time $= 4N_e$

average time between substitutions $\frac{1}{\mu}$

4.3 Rates corrected for multiple hits

Corrected replacement rate

Poisson process model for one amino acid site

replacement number $X \in \text{Pois}(\Lambda u)$ during time u

no reverse mutations for amino acids (20 letters)

Proportion of differences per site

$D = \frac{1}{L}(1_{\{X_1>0\}} + \ldots + 1_{\{X_L>0\}})$

$\text{E}(D) = 1 - e^{-2t\Lambda}$, $\text{Var}(D) = \frac{1}{L}(1 - e^{-2t\Lambda})e^{-2t\Lambda}$

Method of moments estimate: $D = 1 - e^{-2t\hat{\Lambda}}$ implies

Corrected replacement rate $\hat{\Lambda} = -\frac{\ln(1-D)}{2t}$

Estimated K: $\hat{K} = -\ln(1-D)$, $s_{\hat{K}} = \sqrt{\frac{D}{L(1-D)}}$

saturated $D = 1$ gives $\hat{K} = \infty$

Ex 2: bacterial gene

$\hat{K} = 0.1053$, $s_{\hat{K}} = 0.0745$, $\hat{\Lambda} = 0.66 \cdot 10^{-9}$

Markov Chain models

MC is a stochastic model assuming that

given the current state future is independent of past

Transition rates

	To A	To C	To G	To T
From A	—	r_{AC}	r_{AG}	r_{AT}
From C	r_{CA}	—	r_{CG}	r_{CT}
From G	r_{GA}	r_{GC}	—	r_{GT}
From T	r_{TA}	r_{TC}	r_{TG}	—

Equilibrium base composition

$F = (\pi_{\mathrm{A}}, \pi_{\mathrm{C}}, \pi_{\mathrm{G}}, \pi_{\mathrm{T}})$ with $\pi_{\mathrm{A}} + \pi_{\mathrm{C}} + \pi_{\mathrm{G}} + \pi_{\mathrm{T}} = 1$

Substitution rate

$$\lambda = \pi_{\mathrm{A}}(r_{\mathrm{AC}} + r_{\mathrm{AG}} + r_{\mathrm{AT}}) + \pi_{\mathrm{C}}(r_{\mathrm{CA}} + r_{\mathrm{CG}} + r_{\mathrm{CT}})$$
$$+\pi_{\mathrm{G}}(r_{\mathrm{GA}} + r_{\mathrm{GC}} + r_{\mathrm{GT}}) + \pi_{\mathrm{T}}(r_{\mathrm{TA}} + r_{\mathrm{TC}} + r_{\mathrm{TG}})$$

Jukes-Cantor model

JC	A	T	C	G
A	—	α	α	α
T	α	—	α	α
C	α	α	—	α
G	α	α	α	—

$F = (0.25, 0.25, 0.25, 0.25)$

$\lambda = 3\alpha$

JC genetic distance corrected for multiple changes

$\hat{k} = \frac{3}{4}\ln(\frac{3}{3-4d})$, $s_{\hat{k}} = \frac{\sqrt{d(1-d)}}{(1-\frac{4}{3}d)\sqrt{l}}$

$\hat{k} \approx d$ if d is small

$$\boxed{\text{Corrected substitution rate } \hat{\lambda} = \frac{\hat{k}}{2t}}$$

Saturated $d = \frac{3}{4}$ when $\frac{1}{4}$ of sites match by chance

gives $\hat{k} = \infty$

Ex 2: bacterial gene

$\hat{k} = 0.1674$, $s_{\hat{k}} = 0.0576$, $\hat{\lambda} = 1.05 \cdot 10^{-9}$

Kimura two-parameter model

Transitions are more usual than transversions

transversions: purines $\{A,G\} \longleftrightarrow$ pyrimidines $\{T,C\}$

transitions: $A \longleftrightarrow G$ and $T \longleftrightarrow C$

K2P	A	C	G	T
A	—	β	α	β
C	β	—	β	α
G	α	β	—	β
T	β	α	β	—

$F = (0.25, 0.25, 0.25, 0.25)$

$\lambda = \alpha + 2\beta$

K2P genetic distance

$\hat{k} = -\frac{1}{2}\ln(1 - 2P - Q) - \frac{1}{4}\ln(1 - 2Q)$

$P =$ differences per site due to transitions

$Q = p - P =$ differences per site due to transversions

Ex 2: bacterial gene

4 transitions, $P = \frac{4}{60} = 0.0667$

5 transversions, $Q = \frac{5}{60} = 0.0833$

$\hat{k} = 0.1221 + 0.0456 = 0.1677$, $\hat{\lambda} = 1.05 \cdot 10^{-9}$

Ex 4: transition-transversion ratio

α/β ratio for different sequences:

12S rRNA = 1.75, alpha- and beta-globins = 0.66

pseudo eta-globins = 2.7, mtDNA = 9.0

4.4 Molecular clocks

Molecular clock hypothesis: average rates of

molecular evolution λ, Λ are nearly constant over time

Ex 5: alpha-globin data

Table 8.1, p. 330: differences between alpha-globins

D above the diagonal, $\hat{K}$ below the diagonal

Divergence times: Fig 8.6, p. 329

phylogentic tree based on paleontological data

Molecular clock: data fit a straight line, Fig 8.7, p. 330

regression line slope = $2\hat{\Lambda}$, $\hat{\Lambda} = 0.9 \cdot 10^{-9}$

Ex 6: beta-globin data

Primates, $L = 146$, fosil evidence for 6 paires of species

t MY (x_i)	85	60	42	40	30	15
# differences	25.5	24.0	6.25	6.0	2.5	1.0
D	0.175	0.164	0.043	0.041	0.017	0.007
$\hat{K}$ (y_i)	0.192	0.180	0.044	0.042	0.018	0.007

Least squares estimate of the slope $b = 0.00315$

$\hat{\Lambda} = \frac{b}{2} = 1.58 \cdot 10^{-9}$ replacements per site per year

Coefficient of determination

$r^2 = 86\%$ of variation in Y is explained by X

Variation in clock rates

Different substitution rates

for different genes and different taxonomic groups

Episodic clock: substitution is a Poisson process with randomly changing rate (variance larger than mean)

Ex 7: viral clocks

Fig 8.9, p.334: NS gene of influenza virus

$l = 890$, $\lambda = 1.9 \cdot 10^{-3}$ subst. per site per year

pol gene of HIV: $\lambda = 0.5 \cdot 10^{-3}$ per site per year

divergence time between HIV1 and HIV2 is 200 years

Ex 8: clock retardation

Fig 8.10, p. 335: *Adh* gene of Drosophila
slow-down of substitutions in D. pseudoobscura clade

Generation-time effect

Neutral evolution theory prediction:
species with shorter generation times evolve faster
strong effect observed for syn. subst. in mammals

Fig 8.8, p. 332: weak effect for amino-acid replacements
evolutionary rate for proteins is nearly constant across species if time is measured in years, not generations

Explanation by negative selection: Λ decreases with N
N is inversely proportional to generation time

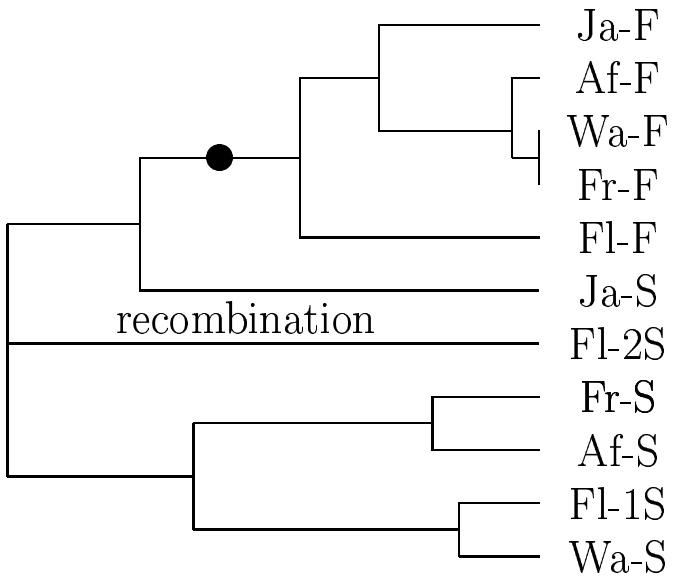
4.5 Gene genealogy and coalescent

Gene genealogy =
tree formed by sequences of alleles from a single species

Ex 9: Adh gene in D.melanogaster

Fig 8.15, p. 346: parsimony tree
for eleven *Adh* alleles in *D.melanogaster*
sampled from different geographical regions
two allozymes Fast and Slow

Branch lengths are proportional to nucleotide differences estimated by parsimony algorithm



Coalescent

a simple stochastic model of a gene genealogy

for n chromosomes sampled from a large population

Coalescent models evolution backward in time

diffusion approximation: evolution forward in time

backward simulations more effective in view of RGD

Coalescent is based on WFM with neutral mutations

reproduction and mutation processes are independent

A unit of coalescent time $= 2N$ generations in WFM

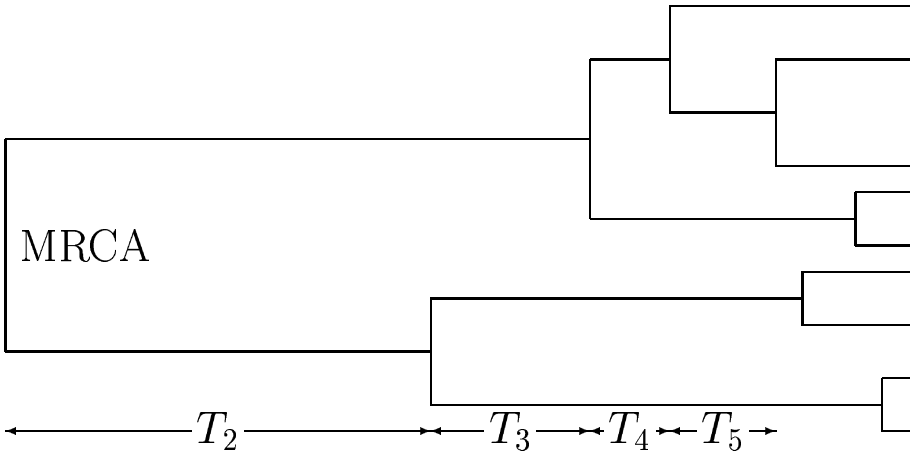
Topology of the coalescent tree

any out of $\binom{n}{2}$ pairs of ancestral lines join first

Coalescent branch lengths: $T_2 \in \text{Exp}(1)$, $T_n \in \text{Exp}\binom{n}{2}$

$\text{E}(T_n) = \frac{2}{n(n-1)}$ more branches - sooner the next merger

$\sigma(T_n) = \frac{2}{n(n-1)}$ huge uncertainty in the tree evolution



Scaled time to the most recent common ancestor

$T_{\text{MRCA}} = T_2 + T_3 + \ldots + T_n$ sum of independent r.v.

$\text{E}(T_{\text{MRCA}}) = 2(1 - \frac{1}{n})$

If $n = 2$, then $T_{\text{MRCA}} = T_2$, $\text{E}(T_2) = 1$, $\text{Var}(T_2) = 1$

If n is large, then $\text{E}(T_{\text{MRCA}}) \approx 2$, $\text{Var}(T_{\text{MRCA}}) \approx 1.16$

Fixation time of a new neutral mutation

is approximately $T_{\text{MRCA}} \times 2N$ with $n = 2N$

the average fixation time $\approx 4N$

Total branch length in the gene tree

$J_n = 2T_2 + 3T_3 + \ldots + nT_n$ sum of independent r.v.

$\text{E}(J_n) = 2a_1$	$a_1 = 1 + \frac{1}{2} + \ldots + \frac{1}{n-1}$
$\text{Var}(J_n) = 4a_2$	$a_2 = 1 + \frac{1}{4} + \ldots + (\frac{1}{n-1})^2$

Total length L_n of the external branches

$\text{E}(L_n) = 2$ is independent of n

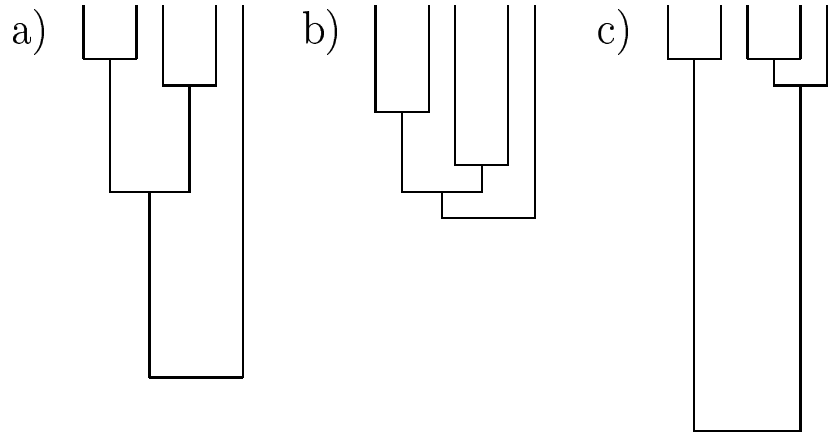
Hypothesis testing using trees

Tree shapes explained by the coalescent theory

a) Theoretical coalescent tree:
   constant population size
   neutral mutations (no selection), no recombination

b) Star-like tree
   growing population size, bottleneck (all loci) or
   positive selection, recent fixation (single locus)

c) Longer branches near the root:
   population subdivision (all loci) or
   balancing selection (single locus)



4.6 Within species molecular polymorphism

Two measures of molecular polymorphism

nucleotide polymorphism $S = \frac{\#(\text{ss})}{l}$, segregating sites

nucleotide diversity $\pi = \frac{\#(\text{pmm})}{\binom{n}{2} \cdot l}$, pairwise mismatches

Alternative way of computing π: $\pi = \frac{n}{n-1}\bar{h}$

average heterozygosity $\bar{h} = \frac{h_1+\ldots+h_l}{l}$

one site heterozygosity $h_i = 1 - \hat{p}_{iA}^2 - \hat{p}_{iC}^2 - \hat{p}_{iG}^2 - \hat{p}_{iT}^2$

Ex 10: Rh3 gene of D.simulans

$n = 5$ aligned sequences of length $l = 500$

16 segregating sites, $S = \frac{16}{500} = 0.032$

One non-synonimous polymorphism at site 142

find what is odd about 142

1	**1**	1	1	1	2	2	2	2	3	3	3	3	4	4	4
3	**4**	6	9	9	0	0	4	4	5	5	7	7	0	1	8
2	**2**	2	2	8	1	7	0	6	1	4	2	5	5	7	3
T	C	T	A	C	C	T	C	C	T	C	G	G	T	T	A
T	C	C	T	A	C	C	T	C	C	T	G	G	T	T	T
C	T	C	C	C	C	C	T	C	T	T	T	G	C	T	A
C	T	C	C	C	C	C	T	T	C	T	G	A	C	T	T
C	T	C	C	C	T	C	T	T	T	T	G	G	C	C	A

$\pi = \frac{79}{5000} = 0.0158$

$\bar{h} = \frac{6.32}{500} = 0.0126$, same $\pi = \frac{5}{4} \cdot 0.0126 = 0.0158$

configuration	(5,0)	(4,1)	(3,2)	(3,1,1)	tot
number of sites	484	9	6	1	500
number of pmm/site	0	4	6	7	
total number of pmm	0	36	36	7	79
h_i	0	0.32	0.48	0.56	
sum of h_i	0	2.88	2.88	0.56	6.32

Infinite-sites model

ISM is a narrower version of IAM assuming that

new mutations occur at sites not previously mutated

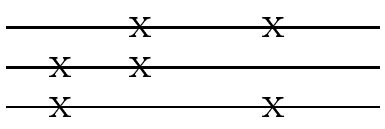
Number of mutations in the gene tree since MRCA

= number of alleles in IAM

= number of segregating sites in ISM

3 alleles possible under IAM

but impossible under ISM



If ISM holds, then easier tree reconstruction

Neutral mutation rate estimation

Consider n aligned sequences of length l assuming ISM

number of segregating sites $l \cdot S \in \text{Bin}(2Nl \cdot J_n, \mu)$

J_n = total branch length in the coalescent

μ = mutation rate per nucleotide site per generation

Two unbiased estimates of θ

$\hat{\theta} = S/a_1$ with $\text{E}(\hat{\theta}) = \theta$ and π with $\text{E}(\pi) = \theta$

$\hat{\theta}$ is consistent, π is inconsistent

$\text{Var}(\hat{\theta}) = \frac{\theta}{la_1} + \frac{a_2\theta^2}{a_1^2}$

$\text{Var}(\pi) = \frac{b_1}{l}\theta + b_2\theta^2$, $b_1 = \frac{n+1}{3(n-1)}$, $b_2 = \frac{2(n^2+n+3)}{9n(n-1)}$

Stochastic variance component

$\lim_{n\to\infty} \text{Var}(\pi) = \frac{\theta}{3l} + \frac{2}{9}\theta^2$

due to sequence dependence by common ancestry

Clustering effect of alleles

coalescent is dominated by T_2, two major clusters

positive covariation of pmm due to few major clusters

new sequences add little information

Ex 11: human effective population size

mtDNA data: 21 humans of diverse origin

868 nucleotide sites with $\pi = 0.0018$

no recombination, high mutation rate

Haploid maternal inheritance implies that

under neutrality π is close to $\theta = 2N_f\mu = N_e\mu$

N_f = effective population size for females

Mammalian mtDNA mutation rate

$5{\cdot}10^{-9}$ to $10{\cdot}10^{-9}$ nucl. subst. per site per year

$\mu = 10^{-7}$ to $2{\cdot}10^{-7}$ subst. per site per generation

human $N_e = \frac{\theta}{\mu} = 9,000$ to 18,000

Fig 8.24, p. 364: star shaped tree, mitochondrial Eve

lived between 180,000 and 360,000 years ago in Africa

4.7 Synonimous and non-synonimous rates

Genetic code is redundant Table 8.2, p. 339

three types of sites: 0 = non-degenerate site

2 = two-fold site and 4 = four-fold site

At a two-fold site $\frac{1}{3}$ of substitutions are synonimous

Effective numbers of sites

$l_s = l_4 + \frac{1}{3} \cdot l_2$ and $l_n = l_0 + \frac{2}{3} \cdot l_2$

total length $l = l_0 + l_2 + l_4 = l_s + l_n$

Fig 8.12, p. 341 and Fig 8.13, p. 342

different substitution rates $\lambda_s = \frac{d_s}{2t}$ and $\lambda_n = \frac{d_n}{2t}$

$d_s = \frac{\text{synonimous changes}}{l_s}$ and $d_n = \frac{\text{nonsynonimous changes}}{l_n}$

Usually $\lambda_s > \lambda_n$ because of deleterious mutations

Fig 8.14, p. 343: mammalian nuclear DNA rates

Neutrality: $\lambda_s = \lambda_n$, positive selection: $\lambda_s < \lambda_n$

Genome averages

Wide variety of unconstrained substitution rates λ_s

plant chloroplast DNA	$1{\cdot}10^{-9}$
mammalian nuclear DNA	$3.5{\cdot}10^{-9}$
plant nuclear DNA	$5{\cdot}10^{-9}$
E.coli and Salmonella enterica bacteria	$5{\cdot}10^{-9}$
Drosophila nuclear DNA	$1.5{\cdot}10^{-8}$
mammalian mitochondrial DNA	$5.7{\cdot}10^{-8}$
HIV-1	$6.6{\cdot}10^{-3}$
Influenza A virus	$1.3{\cdot}10^{-2}$

Ex 2: bacterial gene

Observed differences per site

$d_s = \frac{7}{10+12/3} = 0.5$, $d_n = \frac{2}{38+12{\cdot}2/3} = 0.04$

uncorrected estimates $\tilde{\lambda}_s = 3.1 \cdot 10^{-9}$, $\tilde{\lambda}_n = 0.3 \cdot 10^{-9}$

Positive selection evidence

in a study of 3595 groups of homologous sequences only 17 groups with λ_n/λ_s significantly larger than 1

many of these are sex-related genes (favor speciation)

In some immunoglobulin genes

$\lambda_n/\lambda_s > 1$ in certain regions

overdominant selection for antibody diversity

4.8 Two tests of neutrality

H_0: observed polymorphism is due to selective neutrality of mutations and not due to natural selection

McDonald-Kreitman test

Chi-square test of homogeneity comparing two pairs of numbers of (synonimous, non-synonimous) differences

1. fixed differences between species
2. within species polymorphic sites

Reject the null hypothesis of neutrality

if two distributions are significantly different

Ex 12: G6PD gene in Drosophila

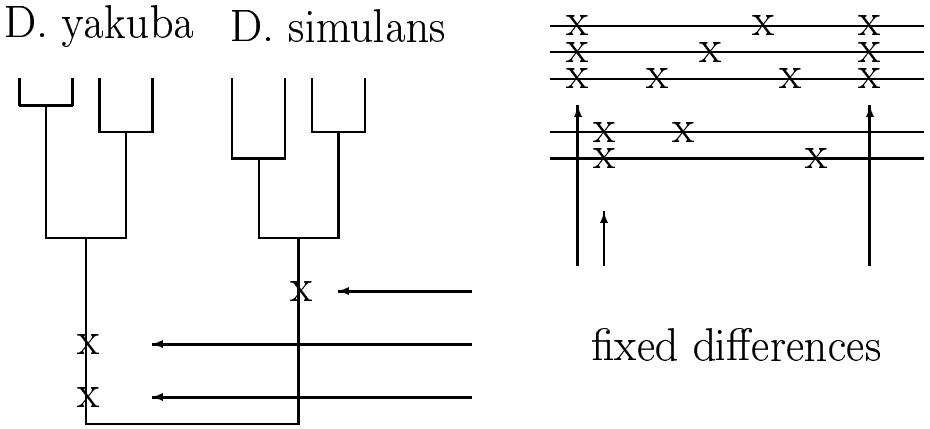
12 alleles in *D.yakuba* and 6 alleles in *D.simulans*

differences	between species	within species	total
synonymous	17(20.3)	29(25.7)	46
nonsynonymous	6(2.7)	0(3.3)	6
total	23	29	52

Excess of nonsynonimous fixed differences

positive selection of advadtageous nonsyn. mutations

$X^2 = 8.6$, df $= 1$, $\sqrt{8.6} = 2.93$, $P = 0.0034$, reject H_0



Tajima test

tests neutrality using polymorphisms in one species

compares two estimates of θ: $\hat{\theta} = S/a_1$ and π

S and π react differently on presence of selection

S examines the number of polymorphic sites

π assesses the site frequences p_A, p_G, p_T, p_C

Very unequal p_A, p_G, p_T, p_C imply smaller π
almost equal p_A, p_G, p_T, p_C imply larger π

Ex 13: configuration and nucletide diversity

$n = 12$, $l = 1$, number of pairs of sequences $\binom{12}{2} = 66$

config	(10,1,1,0)	(9,1,1,1)	(6,3,2,1)	(4,3,3,2)	(3,3,3,3)
#(pmm)	21	30	48	53	54
π	0.318	0.455	0.727	0.803	0.818

Tajima test statistic

Under hypothesis of neutrality $\mathrm{Var}(\pi - \frac{S}{a_1}) = \frac{c_1\theta}{l} + c_2\theta^2$

where $c_1 = b_1 - \frac{1}{a_1}$, $c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}$

$$\boxed{D = \frac{\pi - S/a_1}{\sqrt{e_1 S + e_2 S(S-1/l)}}, \text{ where } e_1 = \frac{c_1}{a_1},\ e_2 = \frac{c_2}{a_1^2 + a_2}}$$

Null distribution of Tajima's D is tabulated by simulation

might be approximated by a Beta distribution

Significant $D > 0$ means almost equal p_A, p_G, p_T, p_C:

either balancing selection (overdominance) or

diversifying selection when rare alleles are favored

Significant $D < 0$ means very unequal p_A, p_G, p_T, p_C:

either selection against rare alleles or

recent bottleneck implying reduced genetic variation

4.9 Recombination and polymorphism

Fig 5.9, p. 182: evolutionary benefit of recombination

Fig 5.10, p. 183: low recomb. rate $\rightarrow$ low polymorphism

Hitchhiking

Fig 8.20, p. 355: hitchhiking (selective sweep) example

advantageous mutation results in reduced

number of segregating sites in tightly linked region

Background selection

reduced diversity at a neutral locus

tightly surrounded by many loci of harmful mutations