Lecture 12. Molecular clocks. Gene genealogy and coalescent. Within species molecular polymorphism.

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

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

Generation-time effect

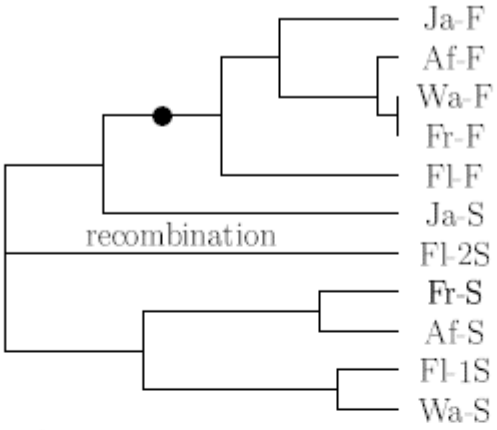
Neutral evolution theory prediction:
species with shorter generation times evolve faster
strong effect observed for syn. subst. in mammals
Fig 8.8, p. 332: weak effect for amino-acid replacements
evolutionary rate for proteins is nearly constant across
species if time is measured in years, not generations
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

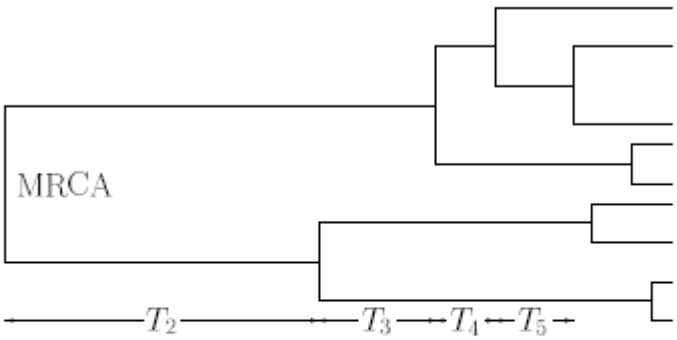
Fig 8.15, p. 346: parsimony tree
for eleven *Adh* alleles in *D.melanogaster*
sampled from different geographical regions
two allozymes Fast and Slow
Branch lengths are proportional to nucleotide
differences estimated by parsimony algorithm



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

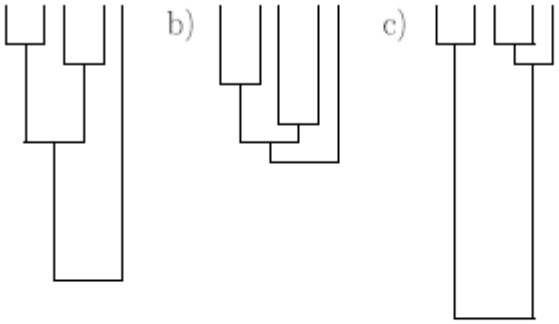
positive selection, recent fixation (single locus)

c) Longer branches near the root:

population subdivision (all loci) or

balancing selection (single locus)

a) b) c)



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: $\pi = \frac{n}{n-1}\bar{h}$

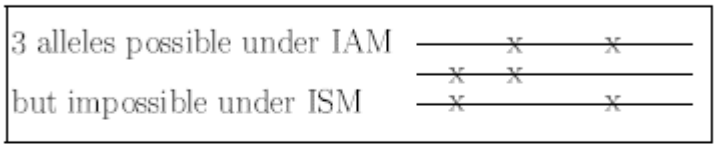
average heterozygosity $\bar{h} = \frac{h_1 + \ldots + h_l}{l}$

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

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



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

Literature:

1. D.L.Hartl, A.G.Clarc. Principle of population genetics. Sinauer Associates, 2007.
2. R.Nielson, M. Statkin. An introduction to population genetics: theory and applications, Sinauer Associates. 2013.