Lecture 11. Theories of molecular evolution. Sequence divergence rates. Rates corrected for multiple hits.

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

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

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

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

Substitution rate

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

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

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.