Chapter 8 – Molecular Evolution

Neutral/Nearly Neutral Theory

Measuring Divergence & Polymorphism

The Molecular "Clock"

Variation in Molecular Rates

Tests for Deviation from Neutral Expectations

Molecular Evolution at Linked Loci/Sites

Sequence Divergence

- simple genetic distance, d = the proportion of sites that differ between two aligned, homologous sequences
- given a constant mutation/substitution rate, d should provide a measure of time since divergence
  - but this is greatly complicated by **multiple hits** (homoplasy)
- given that there are **not** an infinite number of sites in a sequence, how is d expected to change with time?

consider two recently diverged sequences…

ACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGT

consider two recently diverged sequences…

ACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTAAGTACGTACGT

what is the chance that the next substitution obscures the first?

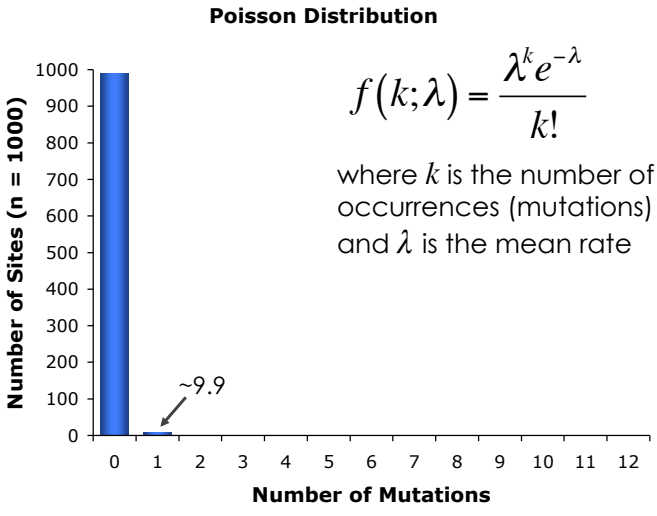
consider two recently diverged sequences…

ACGTACG**C**ACGTACGTACGTACGTACGT

ACGTACGTACGTACGTA**A**GTACGTACGT

now, what is the chance that the next substitution obscures one of the first two?

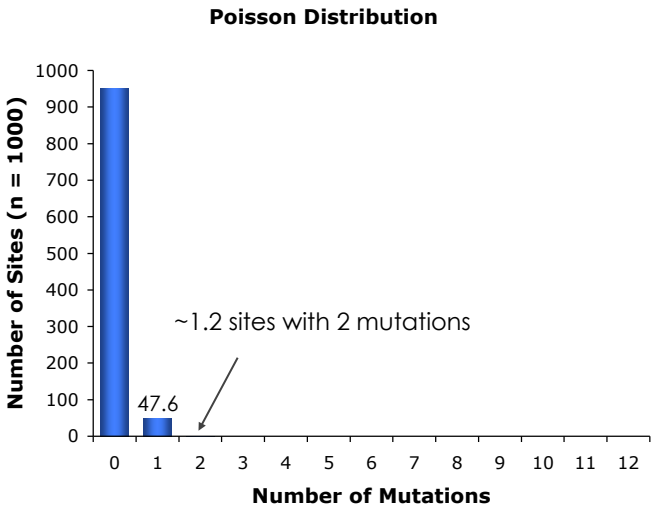
$\lambda = 0.01$ substitutions per site



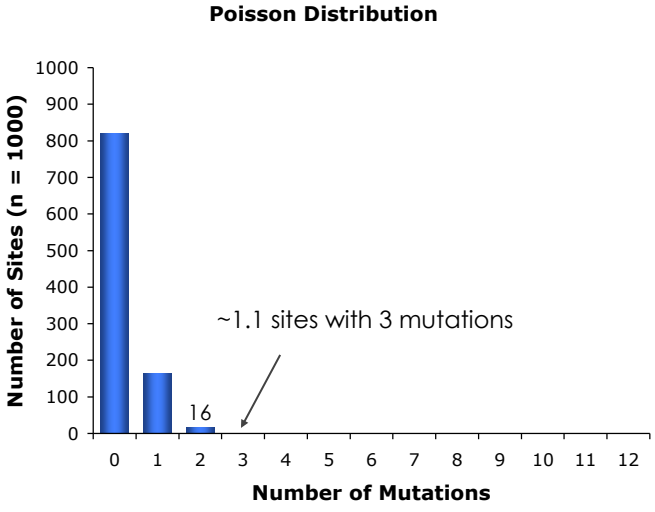
$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where k is the number of occurrences (mutations) and λ is the mean rate

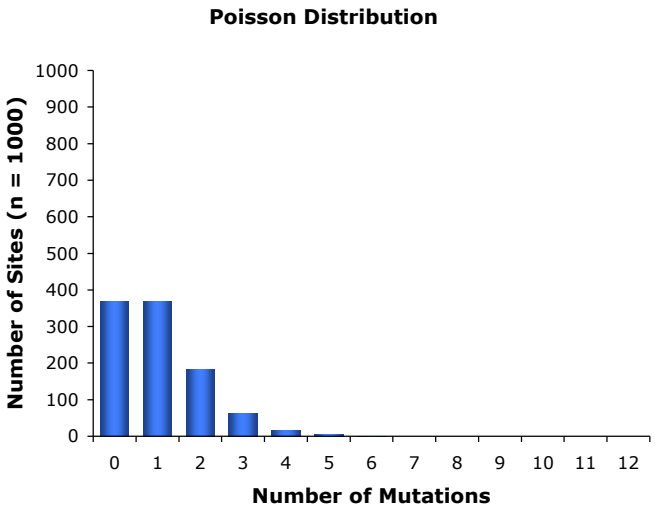
$\lambda = 0.05$ substitutions per site



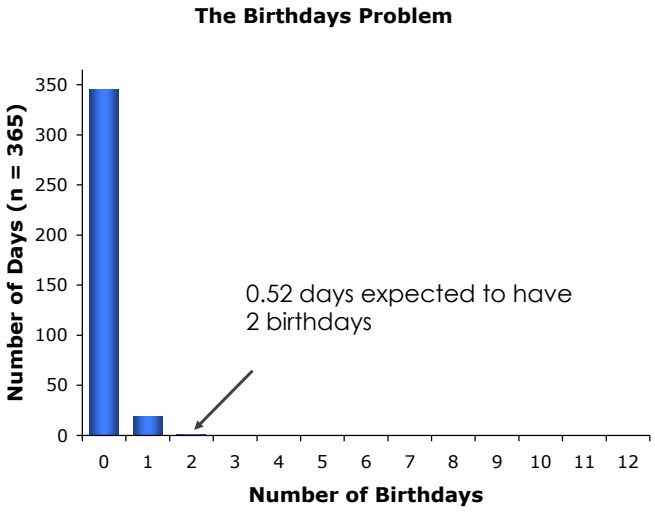
$\lambda = 0.20$ substitutions per site



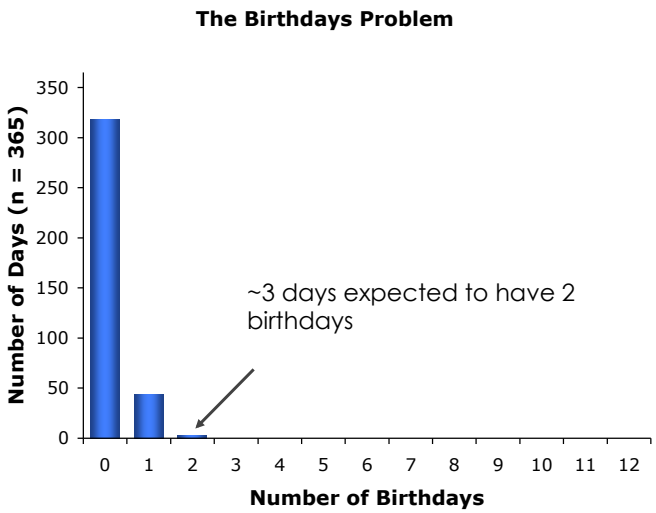
$\lambda = 1.0$ substitutions per site



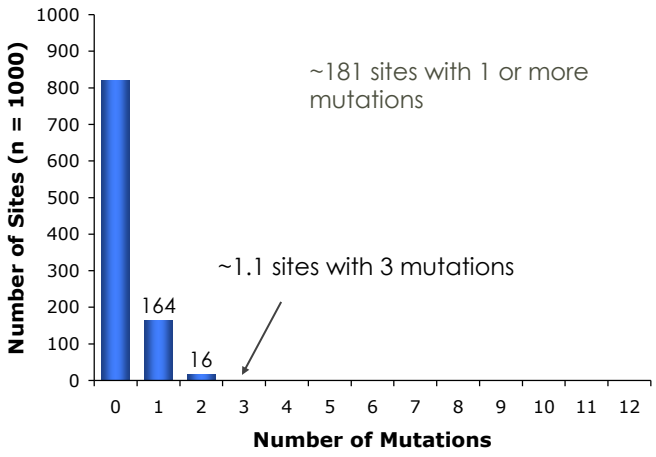
$\lambda = 20/365$ birthdays per day



$\lambda = 50/365$ birthdays per day



$\lambda = 0.20$ substitutions per site



Divergence of DNA Sequences

- even if mutation occurs by a random Poisson process…
  - divergence (genetic distance) depends on changes in both sequences, not just one
  - mutations yield one of four different nucleotides (A, C, G, T)
  - parallel and reverse mutations may result in sequences being identical at a particular position

Jukes-Cantor Distance

$$K = -\frac{3}{4} \times \ln\left(1 - \frac{4}{3}d\right)$$

- where k is an estimate of the number of substitutions that have actually occurred as a function of the observed number of differences d
- assumes a simple model of nucleotide substitution
  - substitutions are equally likely at all sites
  - any nucleotide is equally likely to be substituted for any other nucleotide
  - the four nucleotides occur at equal frequency

Derivation of Jukes-Cantor Distance

- probability that a given site is an A

$$P_{A(t+1)} = (1-3\alpha)P_{A(t)} + \alpha(1-P_{A(t)})$$

  - where α is the mutation rate between each of the four nucleotides

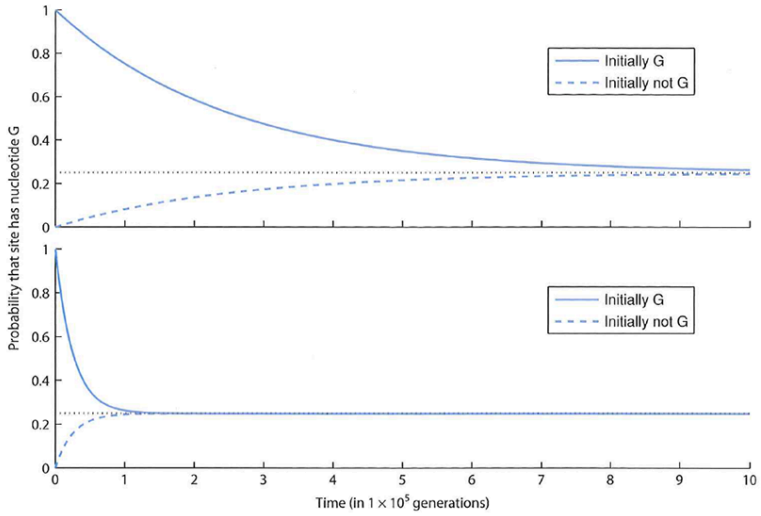


Figure 8.9 The probability that a nucleotide site retains its original base pair under the Jukes–Cantor model of nucleotide substitution. If a nucleotide site originally has a G base, for example, the probability of the same base being present declines steadily over time. If a nucleotide site was initially not a G (it was an A, C, or T), the probability that a G is present at the site increases over time. The probability that a given base is present always converges to 25% because that is the probability of sampling a given base at random if the probability of substitution to each nucleotide is equal. In the top panel $\alpha = 1 \times 10^{-6}$ whereas in the bottom panel $\alpha = 1 \times 10^{-5}$.

Derivation of Jukes-Cantor Distance

- probability that a given site is an A
- solve differential equation for $P_{A(t)}$
- probability that a site remains the same in two lineages
- expected proportion of sites that differ
- rate of change to another nucleotide
- actual number of substitutions per site
- estimate of k based on d

$$P_{A(t+1)} = (1-3\alpha)P_{A(t)} + \alpha(1-P_{A(t)})$$

$$P_{A(t)} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$

$$P_{NN} = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}$$

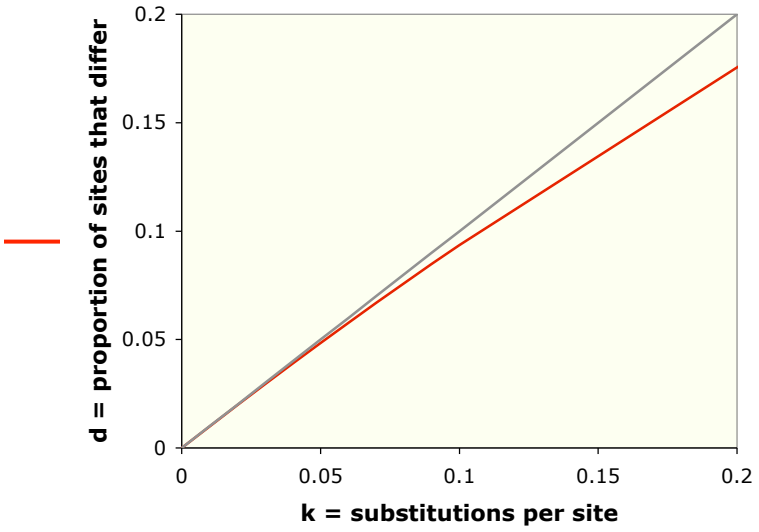
$$d = 1 - P_{NN} = \frac{3}{4}\left(1 - e^{-8\alpha t}\right)$$

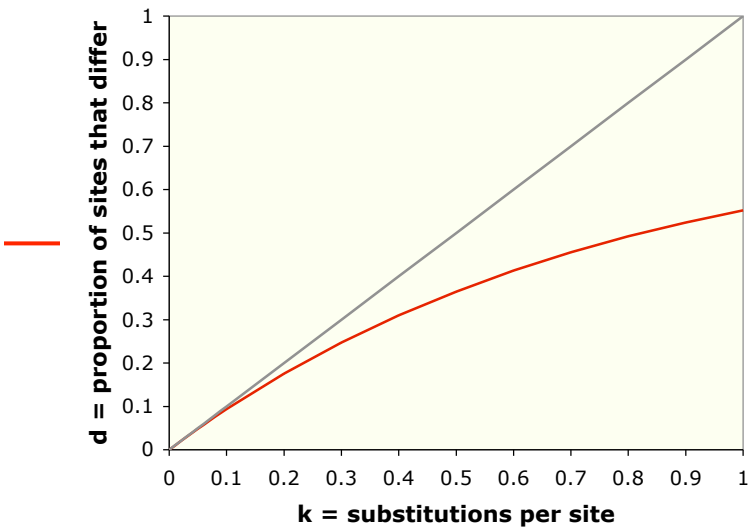
more algebra

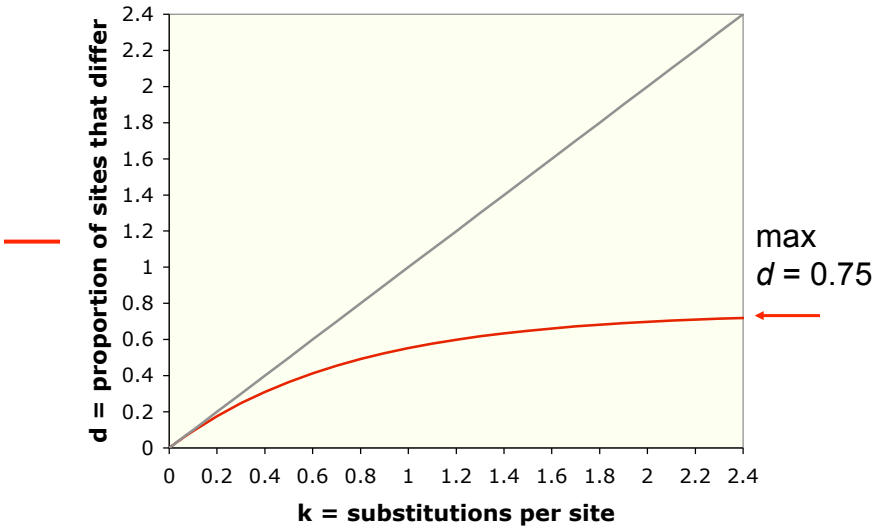
$$\lambda = 3\alpha$$

$$k = 2\lambda t$$

$$K = -\frac{3}{4} \times \ln\left(1 - \frac{4}{3}d\right)$$




d = proportion of sites that differ
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
k = substitutions per site


d = proportion of sites that differ
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2 2.2 2.4
k = substitutions per site
max
d = 0.75

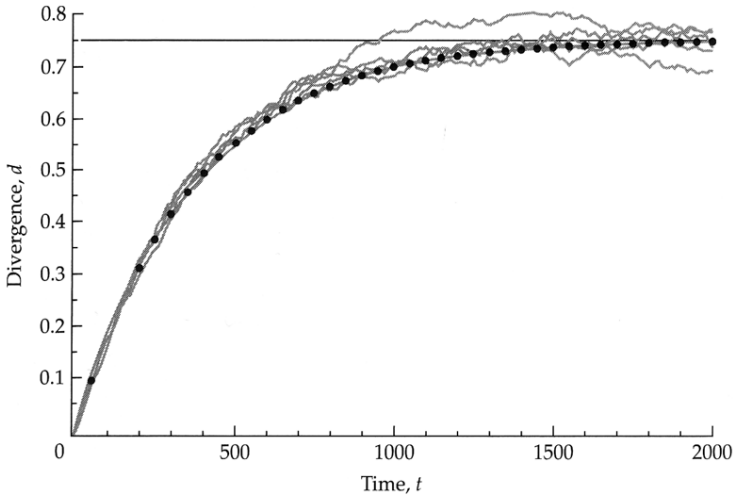
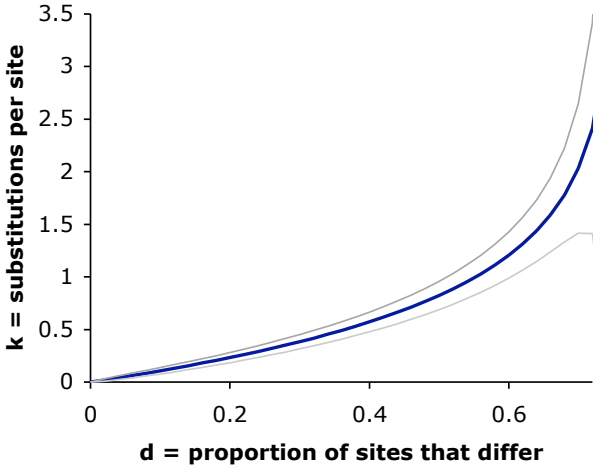


FIGURE 7.5 Simulations of the substitution process for nucleotide sequences show that the sequence divergence saturates at $d = 0.75$ assuming that A, G, C, and T are equally abundant. The jagged lines are numerical simulations of a sequence of length 1000, and the dots give the prediction under the Jukes-Cantor model.

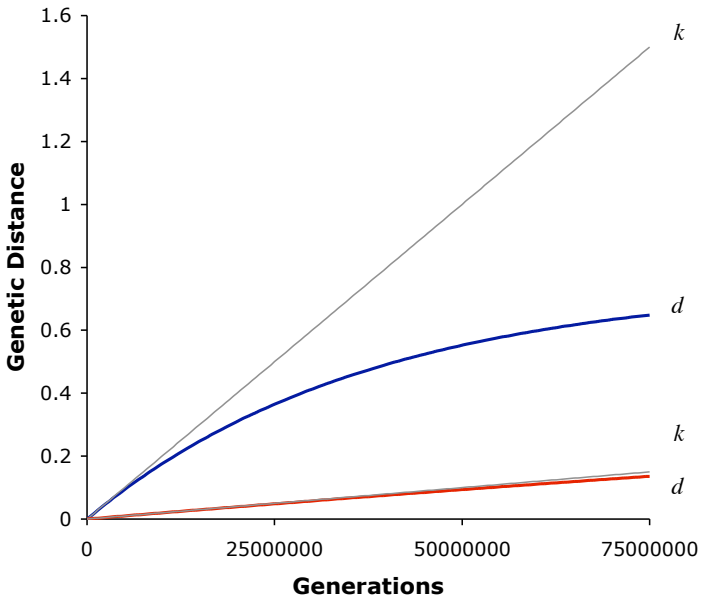
variance in estimate of k increases greatly as d approaches 0.75

$$Var(\hat{k}) = \frac{\hat{d}(1-\hat{d})}{L\left(1-\frac{4}{3}\hat{d}\right)}$$



$k \pm 2SD$
(for L = 500)

Jukes-Cantor equation undefined for $d \geq 0.75$



10^{-8} (e.g., mtDNA)

10^{-9} (e.g., nuclear genes)

*mutation rate in substitutions / site / generation