Molecular Evolution
Answering question 2

The degree of divergence between 2 sequences is the Hamming distance (the edit distance /length of the sequence)

Variables

- Number of mutations (K)
- Rate of mutation (α)
- Time elapsed since divergence (T)

Remember from Junior High Algebra:
time x rate = distance

When divergence is neither too recent nor too remote in time*:

Distance

Rate

$$\alpha = \frac{K}{2T}$$

Time

from paleontological data or by inference from other paleontological data

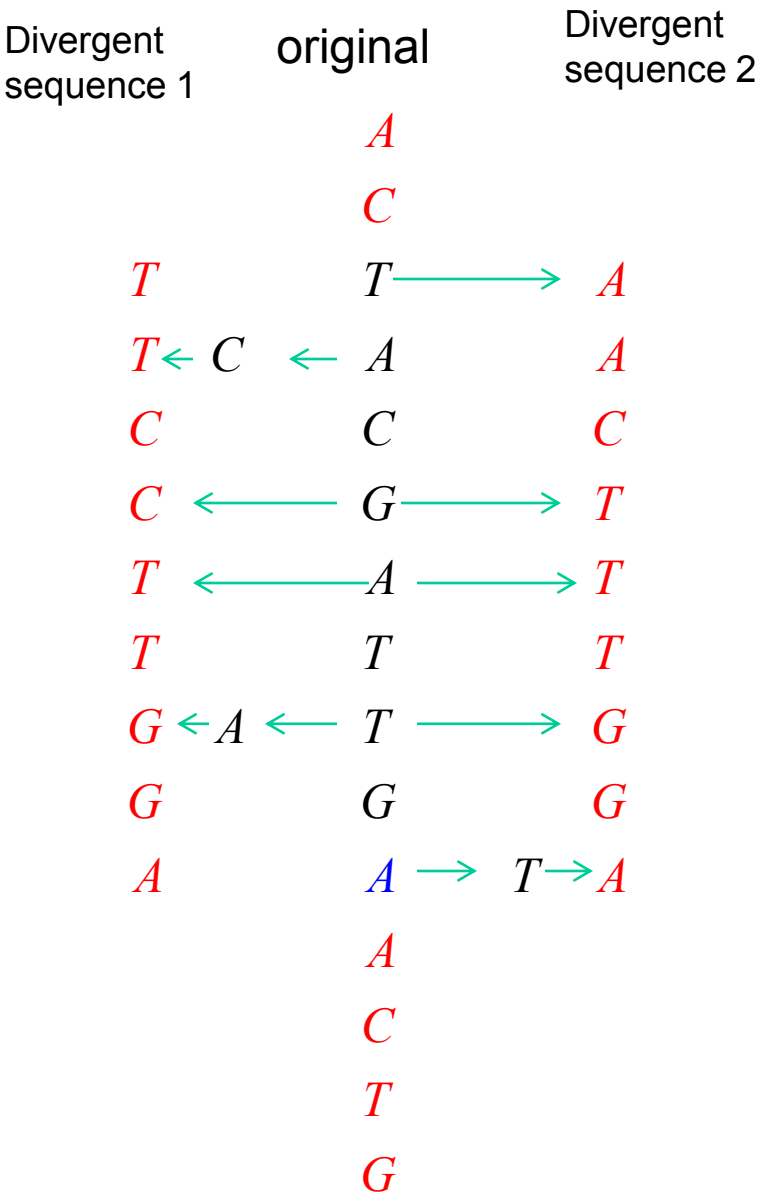
*Polymorphism prior to divergence in very close species
Increased probability of same site multiple substitutions in remote species

Divergent Sequences
Number of Mutations (Evolutionary Distance)

- We really don't know how many mutations have occurred in divergent sequences
  - There can be additional mutations of the same site in one sequence
  - The same site can mutate in both sequences
  - The same site in both sequences can mutate to the same base and appear never to have diverged

Divergent sequences-

Some possible mutation schemes*



*from Grauer and Li

Jukes and Cantor Mutation Model

- If a sequence exists over *t*, the probability of the base, say, A, at any given site being the same is pAA_t
- The joint probability that two (divergent) sequences having the *same* base at the same site is $p(A_0A_t)$for seq1 × $p(A_0A_t)$ for seq 1, or $p^2A_0A_t$
- Likewise, the probability that two (divergent) sequences having a *different* base at the same site is p^2AC_t or p^2AG_t or p^2AT_t
- The total probability is

$$p_{total} = p^2 A_0 A_t + p^2 A_0 C_t + p^2 A_0 G_t + p^2 A_0 T_t$$

Recalling that, for the Jukes and Cantor model,

$$pA_0A_t = \frac{1}{4} + \left(\frac{3}{4}\right)e^{-4\alpha t}$$

And, having just established that

$$p_{total} = p^2A_0A_t + p^2A_0C_t + p^2A_0G_t + p^2A_0T_t$$

we determine that

$$p_{total} = \frac{1}{4} + \left(\frac{3}{4}\right)(e^{-4\alpha t})^2 = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}$$

Now, p_{total} is the probability that we end up with the same nucleotide as we started with, after *t*. For our investigation of divergent sequences, we are really looking for the probability that the nucleotide in a given site would be different after *t*.

That probability is, of course $p_{different}=1-p_{total,}$ or

$$p_{different} = \frac{3}{4}(1 - e^{-8\alpha t})$$

By rewriting

$$p_{different} = \frac{3}{4}(1 - e^{-8\alpha t})$$

we get

$$-8\alpha t = -\ln(1 - \frac{4}{3} p_{different})$$

But we cannot estimate α. We do know, however, that $3\alpha t$ is the rate of substitutions per site .

Let K represent the number of substitutions per site since the sequences diverged. For the Jukes-Cantor model,

$$Arbitrarily, set\ K = 2(3\alpha t) = 6\alpha t\ \ or\ K = -\tfrac{4}{3}6\alpha t$$

$$Substituting\ K\ into\ the\ expression\quad -8\alpha t = -\ln(1-\frac{4}{3}p)\ we\ get$$

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

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

Estimating Evolutionary Distance

K is a proxy evolutionary distance. In the final analysis, α will need to be calibrated, most likely by biological observation

$$dist \approx -\frac{3}{4}\ln(1-\frac{4}{3}p)$$

$$where\ p = fraction\ of\ changed\ nucleotides$$

$$or\ p = \left(\frac{\#\ of\ changes}{length\ of\ sequence}\right)$$

Hamming distance is sometimes defined as the number of changes (same as edit distance) and sometimes as the number of changes/sequence length. Here p is the Hamming distance

EXAMPLE: Consider these two sequences

$A \quad T \quad C \quad G \quad A \quad G \quad C \quad A$

$A \quad A \quad C \quad G \quad A \quad C \quad C \quad A$

The edit distance is 2.

p is 2/8 = .25

Dist= -0.75ln[1-4/3(0.25)]

= 0.30035

When diverging sequences are far apart, distance K becomes unreliable because of sites involved more than once

Substitution rates

- Coding DNA
  - Synonymous substitutions: same AA
  - Nonsynonymous substitutions: different AA
- Non coding DNA
  - Data from UTRs, else scant data

Protein Coding

Synonymous and Nonsynonymous substitutions

- A #1 or #2 position can influence whether #3 will make a synonymous substitution

- Transitions are more frequently synonymous than transversions

All of which make the models significantly more complicated

Codons

- 4-fold degeneracy: any nucleotide in the 3rd position specifies the same AA
  - gly: GGA,GGC,GGG,GGU
- 2-fold degeneracy: two nucleotides in the 3rd position specifiy the same AA
  - glutamic acid: GAA,GAG
    - Only transversions are nonsynonymous
- Special case: 3 nucleotides code for the same AA
  - ileu: AUA,AUC,AUU
- 3 AAs (ser,leu,arg) have 6 codons
- 2AAs (met (AUG) and try (UGG) have only 1 codon

Type of substitution *vis à vis* rate of substitution* (in substitutions/billion yrs)

	Non degenerate	Twofold degenerate	Fourfold degenerate
Transition	0.40	1.86	2.24
Transversion	0.38	0.38	1.47

*Table from Grauer and Li

Rates

Coding DNA

Non-synonymous

actin α 0 substitutions /site /year

γ interferon 3.1×10^{-9} substitutions /site /year

Synonymous up to 25 x higher rate

Substitution rates within genes

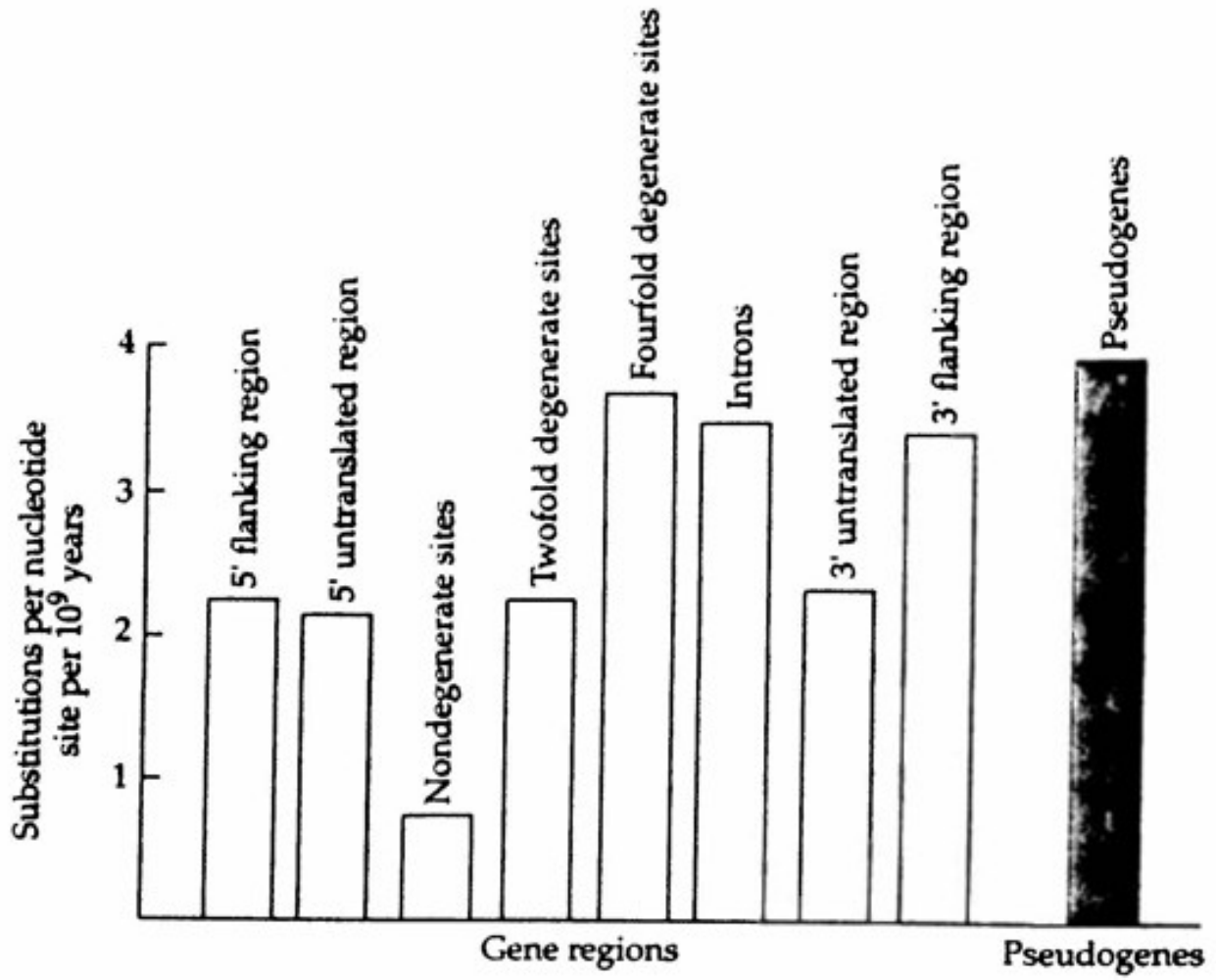


FIGURE 4.3 Average rates of substitution in different parts of genes (white) and i pseudogenes (gray). From Li (1997).

Mutation Rates

Possibly explained by

- Mutational input
- Genetic drift of neutral alleles
- Purifying selection against deleterious alleles (selectional constraint)

from Grauer and Li

But what about positive selection?

If Darwinian positive selection, then $K_{nonsynonymous} > K_{synonymous}$

BUT

Statistical analysis does not lead to that conclusion

MOLECULAR CLOCK CONCEPT*

The assumption: Mutations occur at a fixed rate (α) across time

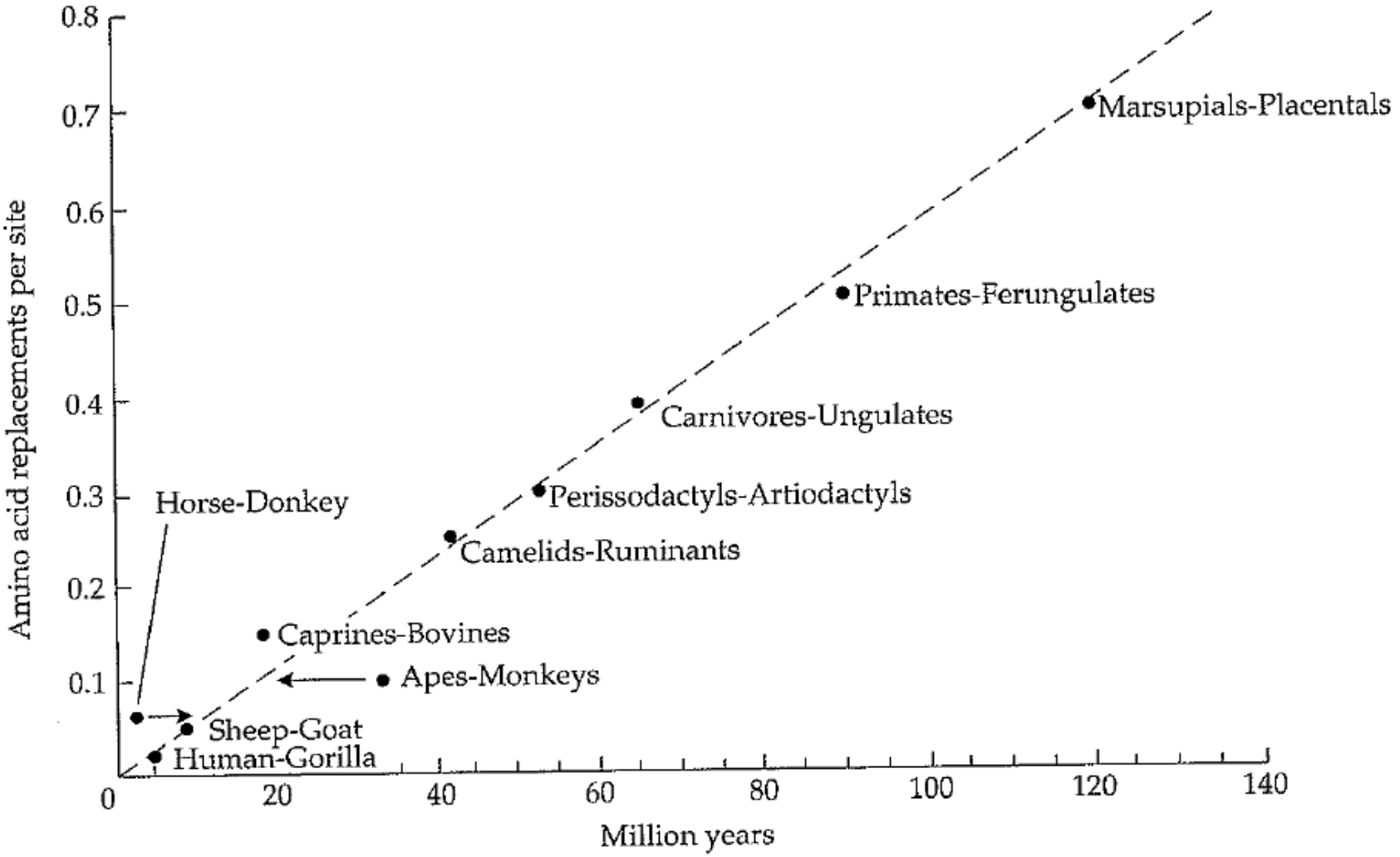
A theory, unproven. But, if indeed there is a molecular clock, then our formula

$$\alpha = \frac{K}{2T}$$

can be used when K is known but there are no paleontological data for T

*Important in phylogeny determinations

Molecular Clock in Action



Taken from Grauer and Li, modified from Langley and Fitch, 1974 Mol Evol 3 161-177 [4]