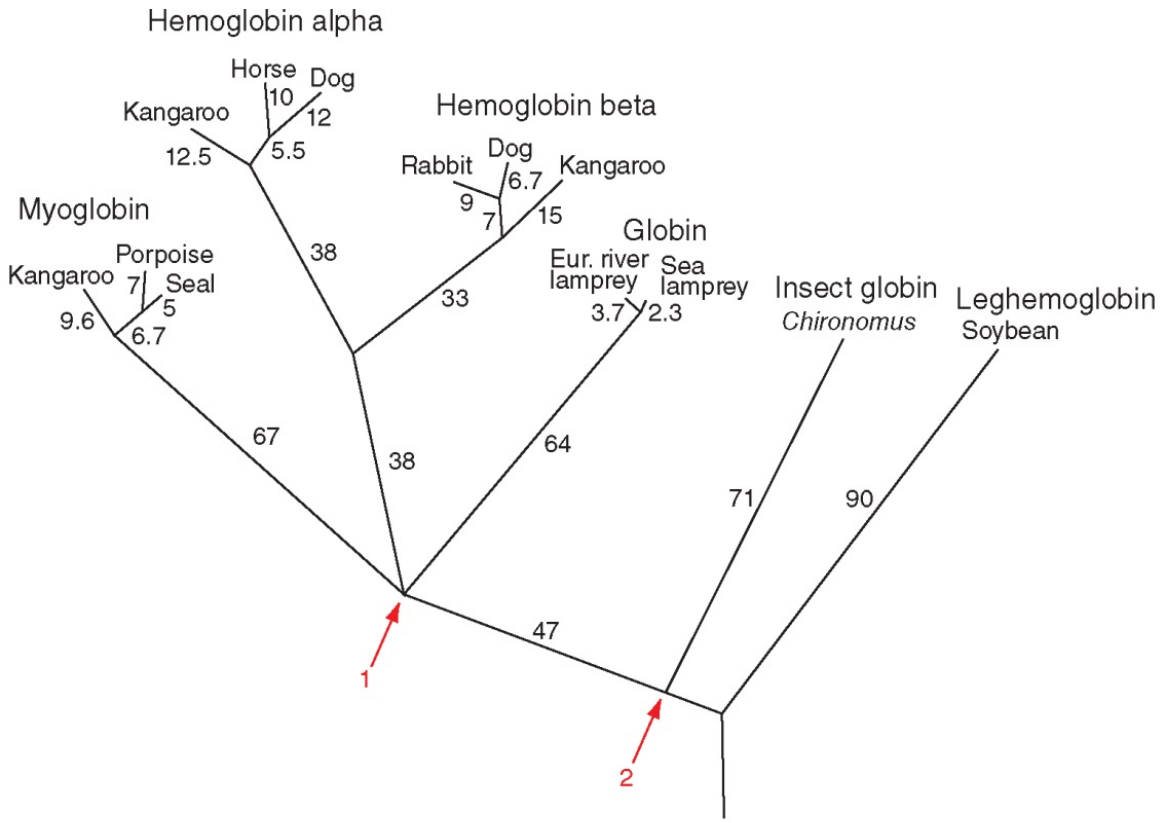
Molecular Evolution

Molecular Evolution

- How and when were genes and proteins created ? How “old” is a gene ? How can we calculate the “age” of a gene ?
- How did the gene evolve to the present form ? What selective forces (if any) influence the evolution of a gene sequence and expression ? Are these changes in sequence adaptive or neutral ?
- How do species evolve? How can evolution of a gene tell us about the evolutionary relationship of species ?

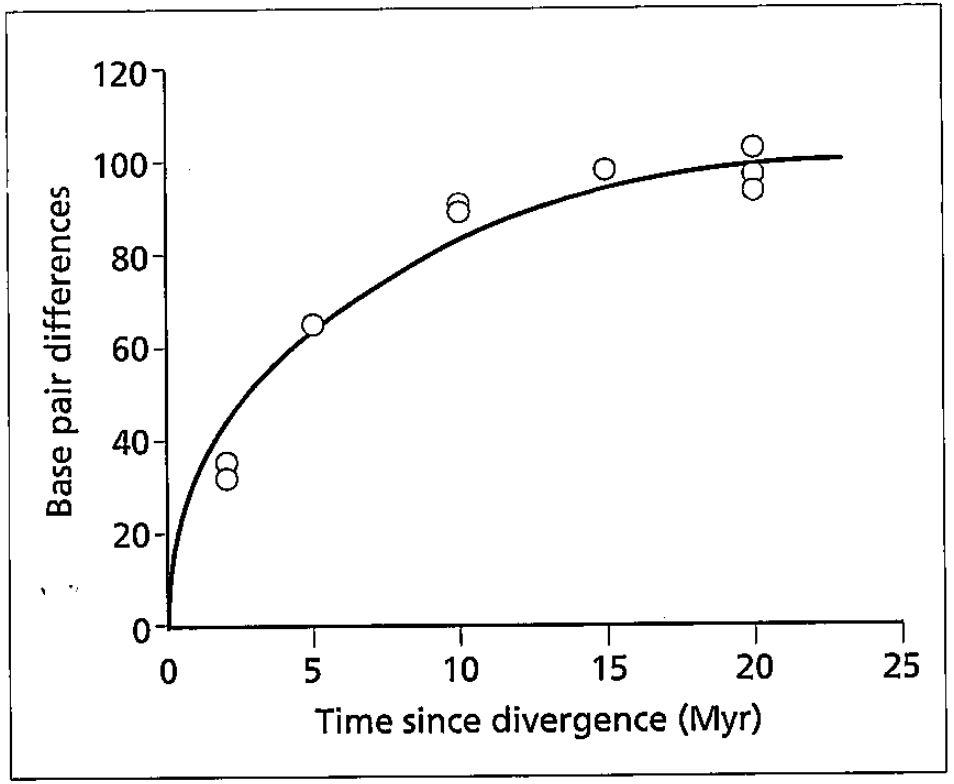
Understanding relationship between homologous sequences

- Complete evolutionary history is depicted as phylogenetic tree
- Tree topology—correctly identify the common ancestors of homologs sequences
- Tree distance –identify the correct relationship in time
- Example: MSA
  - Topology—which sequences should be aligned first
  - Distance—how to weight the sequences when computing alignment score



Measuring evolutionary distance

- How long ago (relatively) did two homologous sequences diverge from a common ancestor
- Simple method: % divergence – 100 - % identity
  - Count up the number of places two sequences differ
- Used by blast: % identity, % similarity
- Easy to compute
- But the **major problem** is that it underestimates divergence after only a moderate amount of change.
- % divergence saturated with time
  - PAM250 represent 80% divergence



Types of nucleotide substitutions

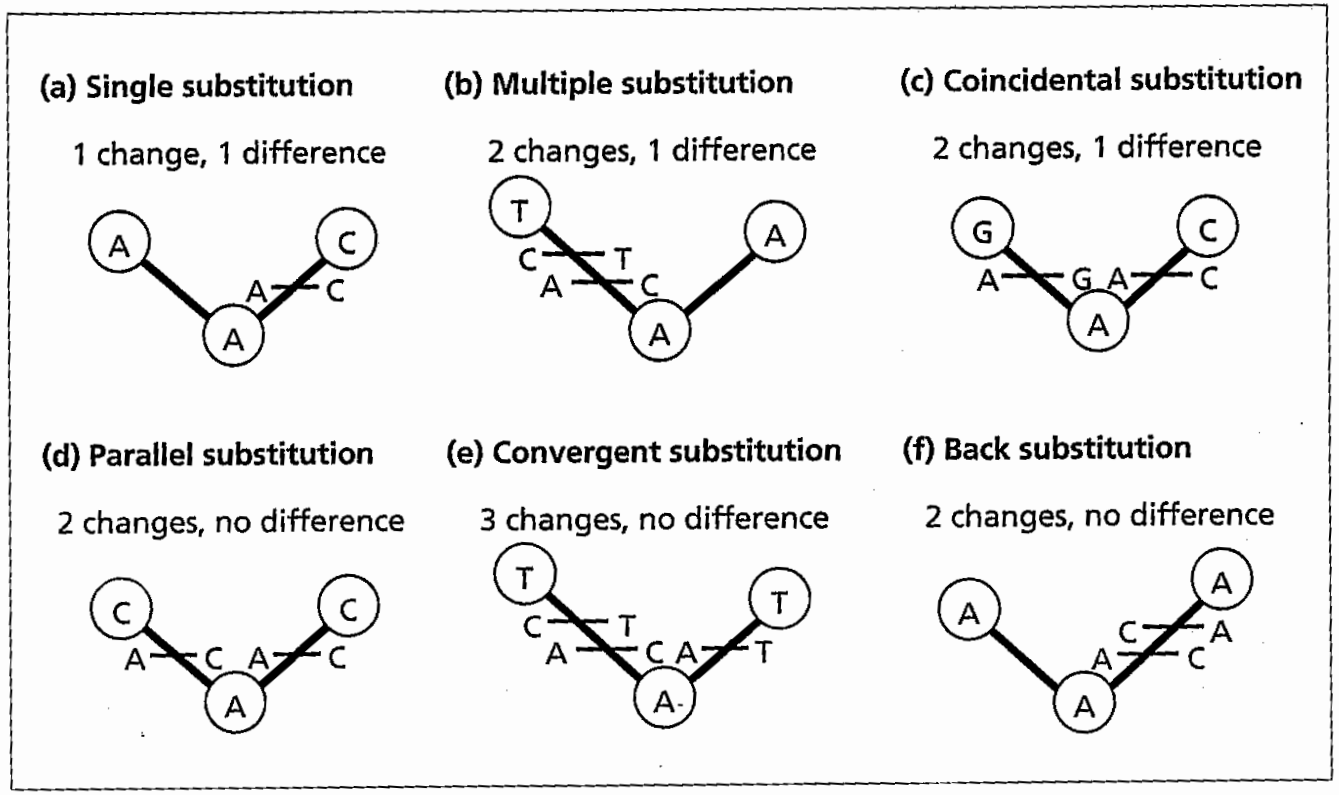
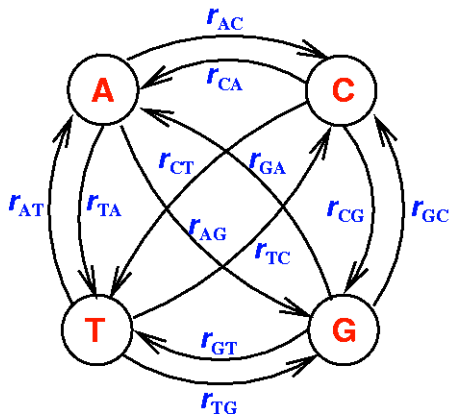


Fig. 5.9 Six kinds of nucleotide substitution. In each case the ancestral nucleotide was A. In all except the case of a single substitution, the number of substitutions that actually occurred is greater than would be counted if we just compared the two descendant sequences. In the lower three cases the nucleotides are identical in both descendant sequences, but this similarity has not been directly inherited from the ancestral sequence. Such similarity is termed 'homoplasious'.

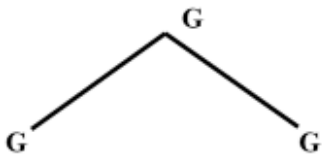
Nucleotide Substitutions Models

- To make functional inferences we typically don't model insertions and deletions, large or small, because of the difficulty in assigning homology.
- We model nucleotide substitutions under the assumptions:
  - They occur as single, independent events
  - They don't affect substitutions at other sites
  - Time scale is much larger than **time to fixation** in a population
- Effective models incorporate the probability of substitutions, and hence account for different kinds of substitutions (multiple, parallel, convergent, reversion)
- Model sequence evolution as a markov process
  - A->A->T->C->A

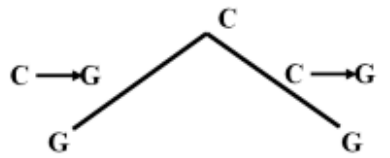


Number of substitutions

- Central question: given two aligned sequences what is the number of substitutions that actually occurred
- Assuming constant substitution rate λ the expected number of substitutions per size is $2\lambda t$
- t is unknown
- Known—observed divergence



. . .G. . .
. . .G. . .



. . .G. . .
. . .G. . .

Jukes-Cantor Model (JC69)

- 1969
- Evolution is described by a single parameter, alpha (α) , the rate of substitution.
- Assumptions:
  - Substitutions among 4 nucleotide types occur with equal probability (rate matrix below)
  - Nucleotides have equal frequency at equilibrium

	A	T	C	G
A	1-3α	α	α	α
T	α	1-3α	α	α
C	α	α	1-3α	α
G	α	α	α	1-3α

Jukes-Cantor Model

What is probability of having nucleotide A $(P_{A(t)})$ at time **t** if we start with A?

- Derive expression P_A using discrete time periods in which the rate is represented as α.
- $P_{A(0)} = 1, P_{C(0)} = 0, \ldots$
- $P_{A(1)} = 1 - 3\alpha, P_{C(1)} = \alpha, \ldots$
- $P_{A(2)} = (1 - 3\alpha)P_{A(1)} + \alpha(1 - P_{A(1)})$
  - at time 1 there was an A, and that A had not changed
  - at time 1 there was not an A and it changed to an A

Jukes-Cantor Model

- Recurrence equation:
  - $P_{A(2)} = (1 - 3\alpha)P_{A(1)} + \alpha(1 - P_{A(1)})$
  - $P_{A(t+1)} = (1 - 3\alpha)P_{A(t)} + \alpha[1 - P_{A(t)}] = (1 - 4\alpha)P_{A(t)} + \alpha$
- What is the change in P_A over time ($\mathbf{\Delta P_A}$)?
- $\Delta P_{A(t)} = P_{A(t+1)} - P_{A(t)}$
- Subtract $P_{A(t)}$ from both sides we get
- $\Delta P_{A(t)} = -4\alpha P_{A(t)} + \alpha$
- $\Delta P_{A(t)}$ is proportional to $P_{A(t)}$

Jukes-Cantor Model

- What is the change in P_A over continuous time?
  - $dP_{A(t)}/dt = -4\alpha P_{A(t)} + \alpha$
- Solve the differential equation
  - $P_{A(t)} = ¼ + (P_{A(0)} – ¼)e^{-4\alpha t}$
- $P_{A(t)} = ¼ + ¾\, e^{-4\alpha t}$ starting at A at t=0
- $P_{A(t)} = ¼ - ¼\, e^{-4\alpha t}$ starting at T,C, or G at t=0
- We can generalize to these equations because all nucleotides are equivalent:
  - $P_{ii(t)} = ¼ + ¾\, e^{-4\alpha t}$
  - $P_{ij(t)} = ¼ - ¼\, e^{-4\alpha t}$

Jukes-Cantor Model

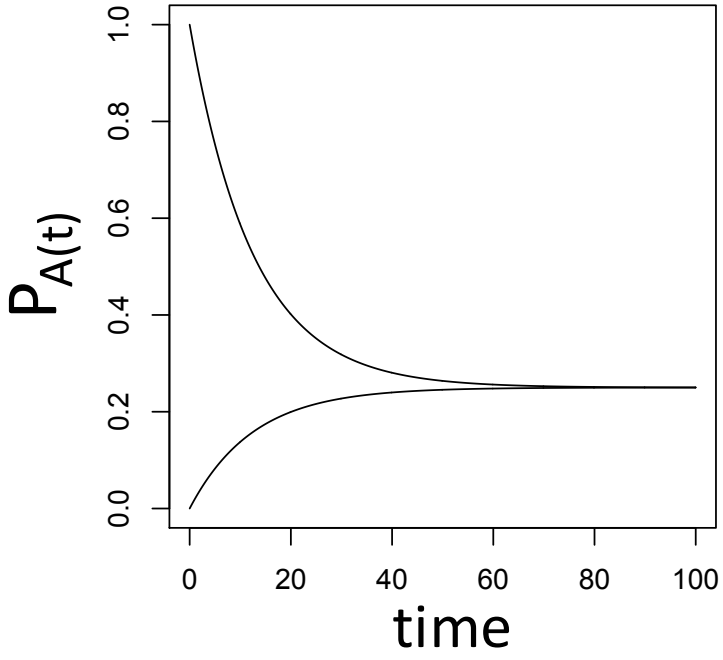
Which equilibrium frequency of *i* is reached at large **t**?

(e.g. *i* = A)

$P_{AA(t)} = ¼ + ¾\ e^{-4\alpha t}$

$P_{jA(t)} = ¼ - ¼\ e^{-4\alpha t}$

We can think of P_i as the frequency of *i* in a long sequence.

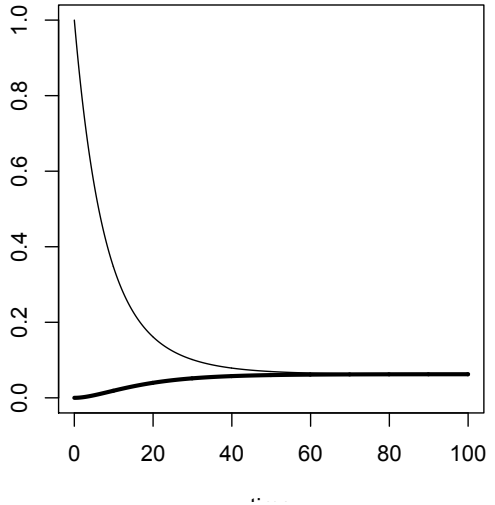


Applying the Jukes-Cantor Model to estimate distance

- Given two sequences evolving independently for a time t what is the probability they both have an A
- $P_{I(t)} = P^2_{AA(t)} + P^2_{AC(t)} + P^2_{AT(t)} + P^2_{AG(t)}$

  both remain A, both change to C, T, or G
- $P_{I(t)} = P^2_{ii(t)} + 3(P^2_{ij(t)})$
- $P_{I(t)} = (¼ + ¾\ e^{-4\alpha t})^2 + 3(¼ - ¼\ e^{-4\alpha t})^2$

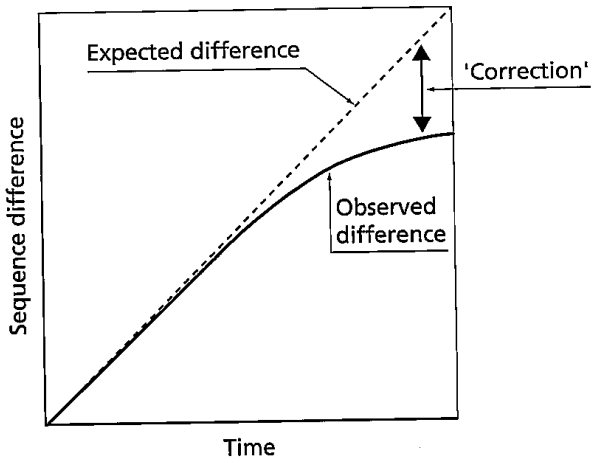
  $= P_{I(t)} = ¼ + ¾\ e^{-8\alpha t}$



1/16

Estimated number of substitutions

- Our goal original goal was to estimate the total number of substitutions since divergence from a common ancestor
- E[sub]=$2\lambda t = 6\alpha t$; $\lambda = 3\alpha$
- Estimate αt from $P_{I(t)} = \frac{1}{4} + \frac{3}{4} e^{-8\alpha t}$ and $P_D = 1 - P_{I(t)}$
  - $\alpha t = -1/8 * \ln(1 - (4/3)P_D)$
- ***E[sub]=¾ ln(1-(4/3)P_D)***--Also known as K_{JC}



Example

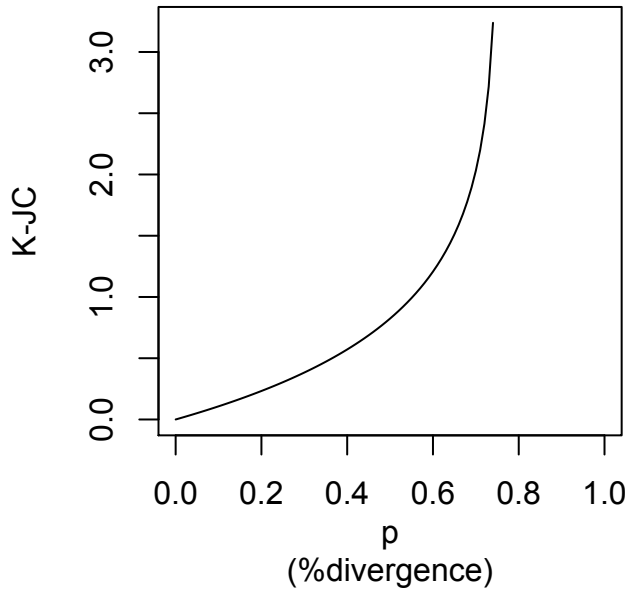
- K_{JC} = function(p) { -0.75 * log(1 - 4*p/3) }

```
1 TATAAAACGGAATGAGGAATAATCGTAATATTAG      34
  ||||||| || |||| |||||||  | |||  |
1 TATAAAATGGTATGATGAATAATATTTATAGAAT      34

Identity:       25/34 (73.5%)
Similarity:     NA/34 (NA%)
Gaps:            0/34 (0.0%)
```

- # divergent bases = total bases – identical bases
  - 34 – 25 = 9

- 9/34 = 0.264705 ← uncorrected % divergence
- K_{JC} (9/34) = 0.326488 ← corrected distance K_{JC}



Rate variation between bases

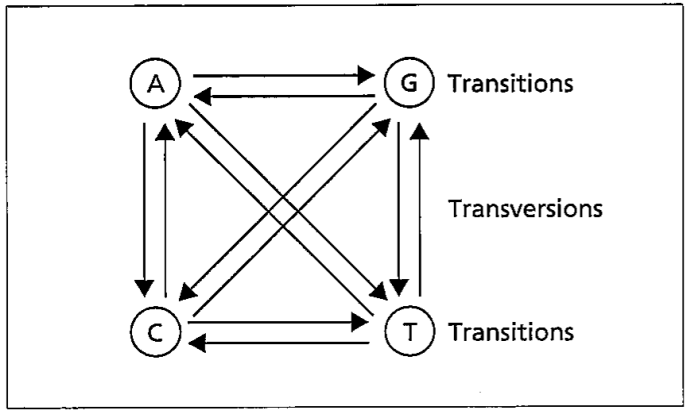
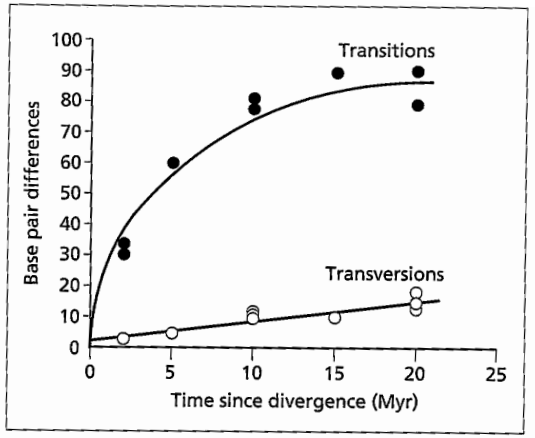
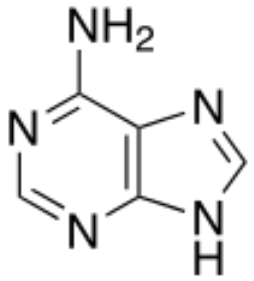


Fig. 5.10 The possible substitutions among the four nucleotides.

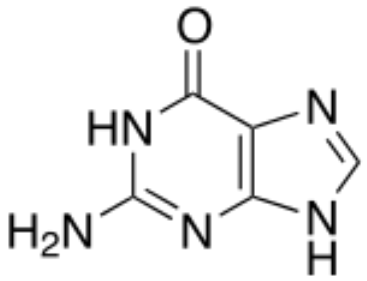


In reality, bases are not equivalent, and rates of change between them are not equal.

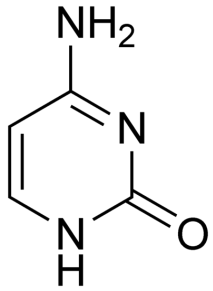
Transitions usually outnumber transversions.



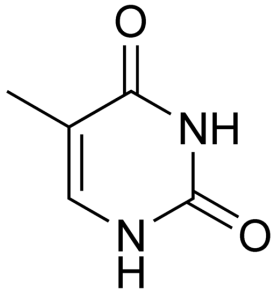
adenine



guanine



cytosine



thymine

Kimura's 2-parameter model (K2P)

- Models transition and transversion rates separately
- Two parameters
  - **α** for transition rate and **β** for transversion rate.
- Assumption:
  - Nucleotides have equal frequency at equilibrium

	A	T	C	G
A	1-α-2β	β	β	α
T	β	1-α-2β	α	β
C	β	α	1-α-2β	β
G	α	β	β	1-α-2β

$f = [0.25,0.25,0.25,0.25]$

Kimura model

- Begin with A: $P_{AA(0)} = 1$
- What is $P_{AA(1)}$?
- $P_{AA(1)} = 1 - \alpha - 2\beta$
- $P_{AA(2)} = (1 - \alpha - 2\beta)P_{AA(1)} + \beta P_{AT(1)} + \beta P_{AC(1)} + \alpha P_{AG(1)}$
- Recursion: $P_{AA(t+1)} = (1 - \alpha - 2\beta)P_{AA(t)} + \beta P_{AT(t)} + \beta P_{AC(t)} + \alpha P_{AG(t)}$

Calculating divergence

P and Q are proportions of divergence due to transitions (P) and transversions (Q) between the 2 sequences

$K_{K2P} = -½ \ln(1 – 2P – Q) – ¼ \ln(1 – 2Q)$

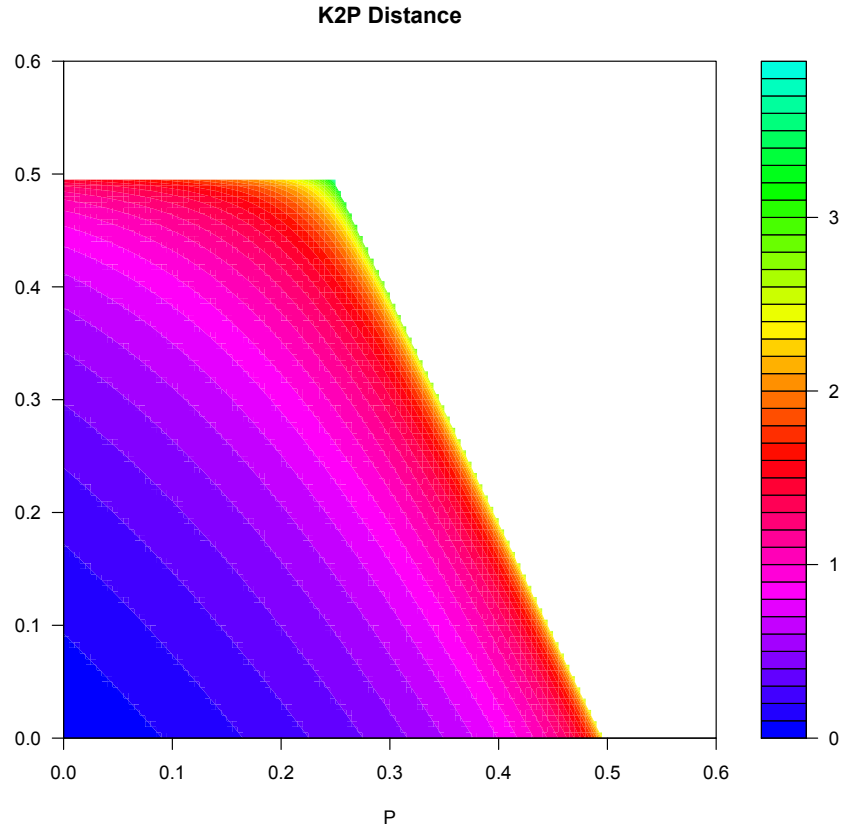
```
k2p = function(p,q)
        { -0.5*log(1-2*p-q) - 0.25*log(1-2*q) }
```

Two 100bp sequences have 20 transitions and 4 transversions between them.

k2p(0.2,0.04) = **0.3107** changes per base pair

31 changes over the entire sequence

The sequences came from a snapdragon (*Antirrhinum*) and a monkey flower (*Mimulus*) whose lineages diverged 76 Mya, yielding a divergence rate of 0.408 substitutions per million years



Kimura K2P model

Estimate the ts/tv ratio

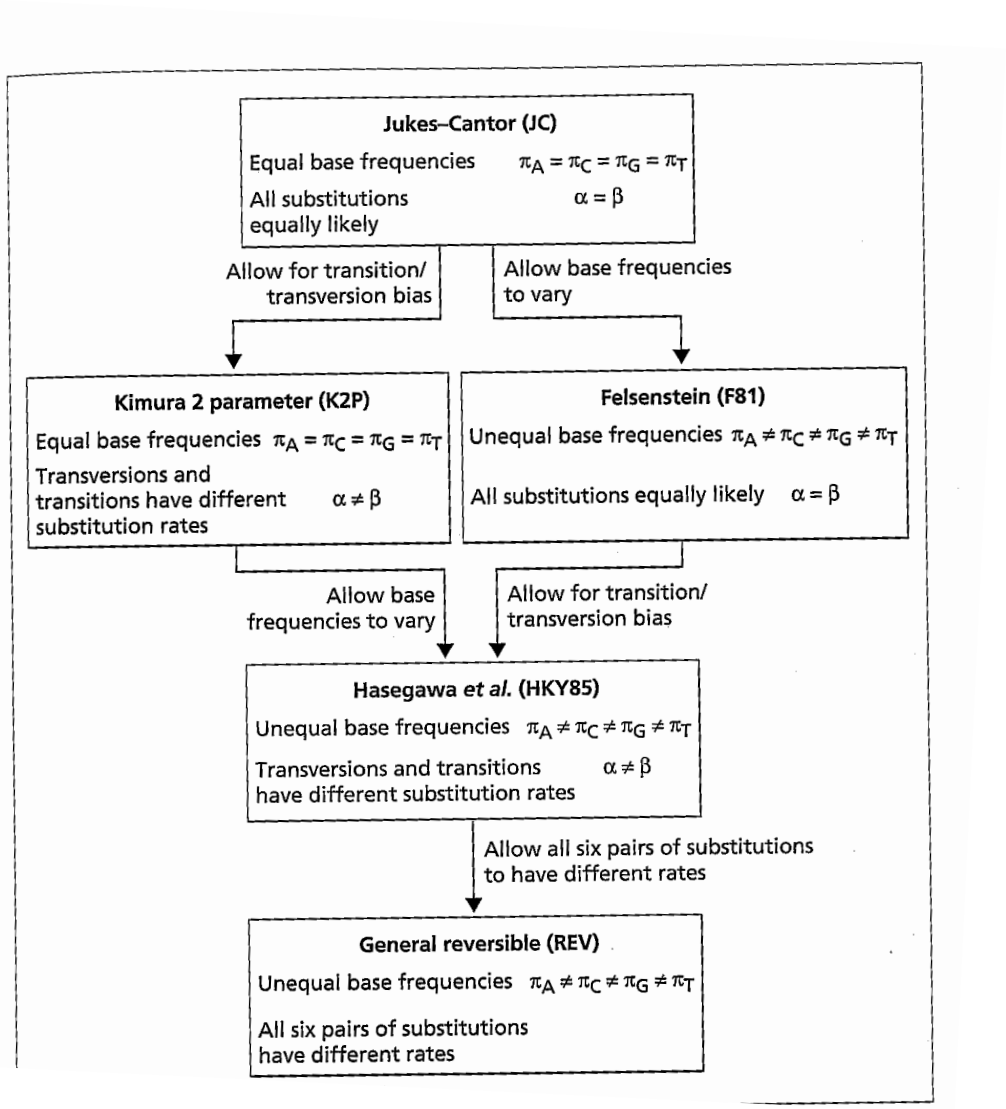
$ts/tv = \alpha/\beta = 2^* \ln(1 - 2p - q) / \ln(1 - 2q) - 1$

Mammals

Nuclear DNA ts/tv ≈ 2

Mitochondrial ts/tv ≈ 15

Nucleotide substitution models



Different substitution models

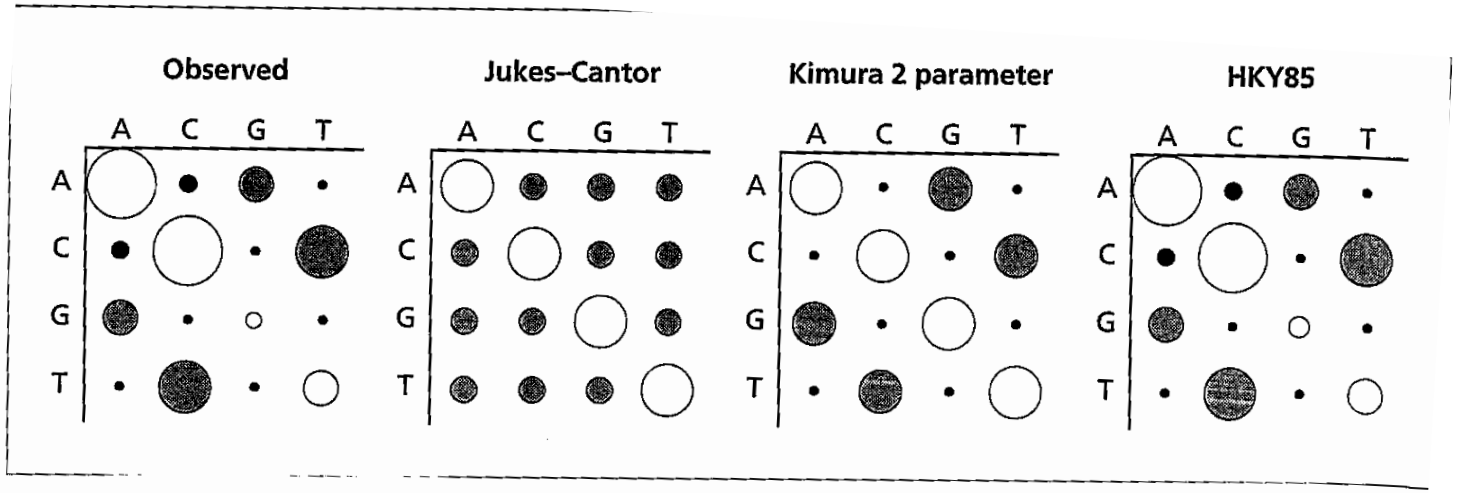
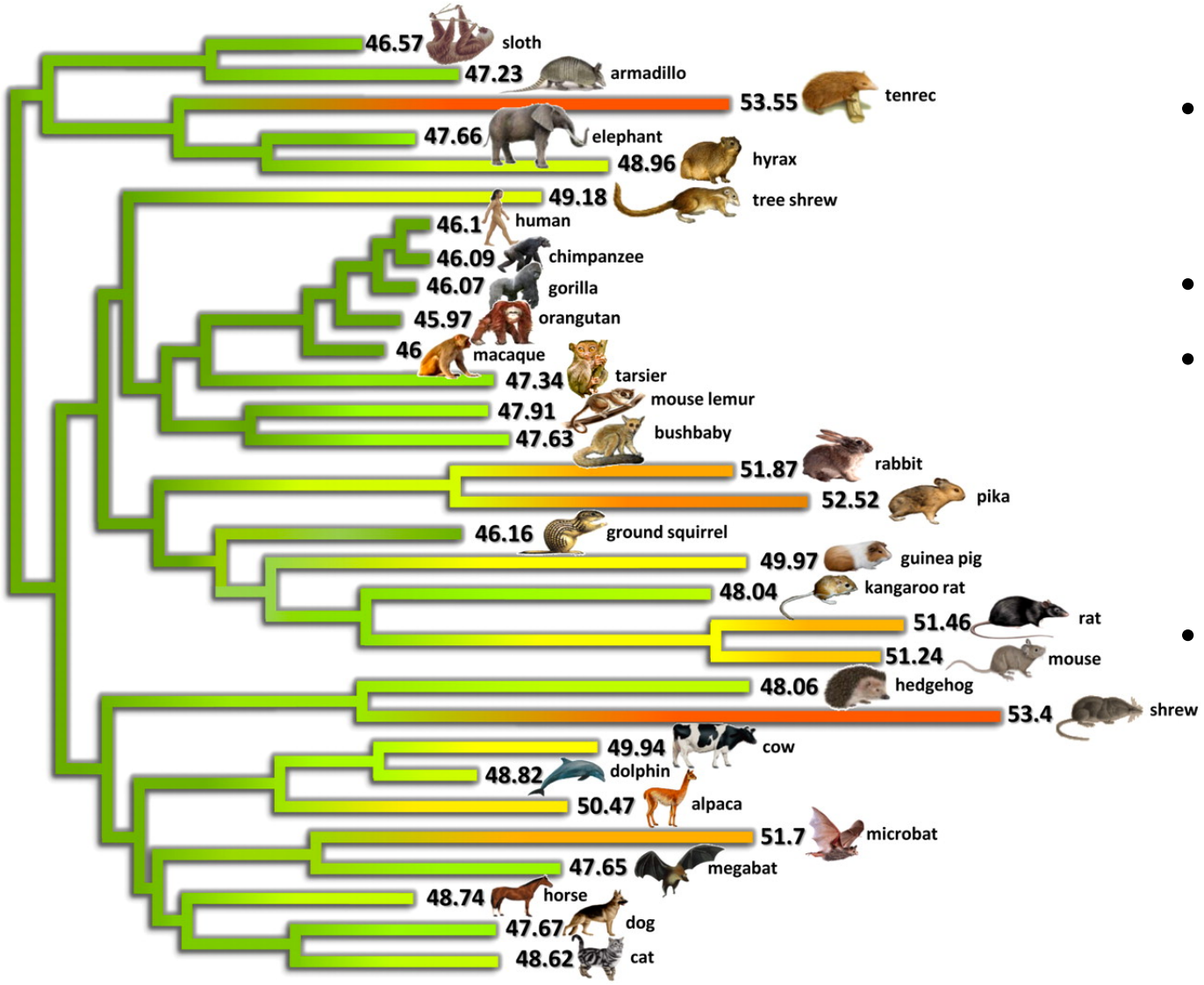


Fig. 5.15 Observed and expected numbers of nucleotide pairs between human and chimpanzee mtDNA sequences for three different models. As the models add parameters they more closely approximate the observed pattern. Data from Tamura (1994).

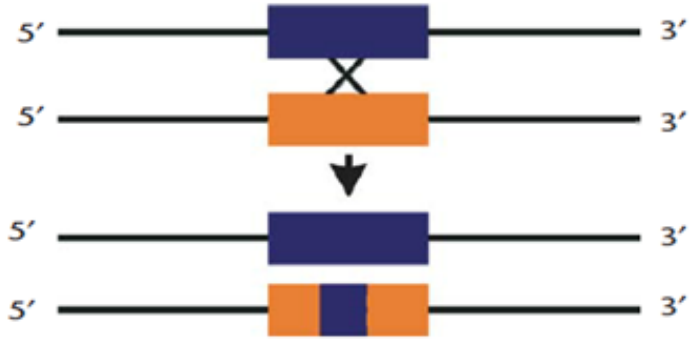
Models have assumptions

- All nucleotide sites have same rate
- The substitution matrix does not change
- Sites change independently.
  - There is no co-evolution or multiple mutation
- This is all true:
  - Sequence changes only due to replication error
  - Errors are randomly propagated –neither advantageous nor deleterious

GC content varies over evolutionary time



- GC content is heterogeneous across the genome at a scale of hundreds of nucleotides
- GC biased gene conversion
- **Gene conversion-i**s the process by which one DNA sequence replaces a homologous sequence such that the sequences become identical after the
- Diploid organism: 2 copies of every locus

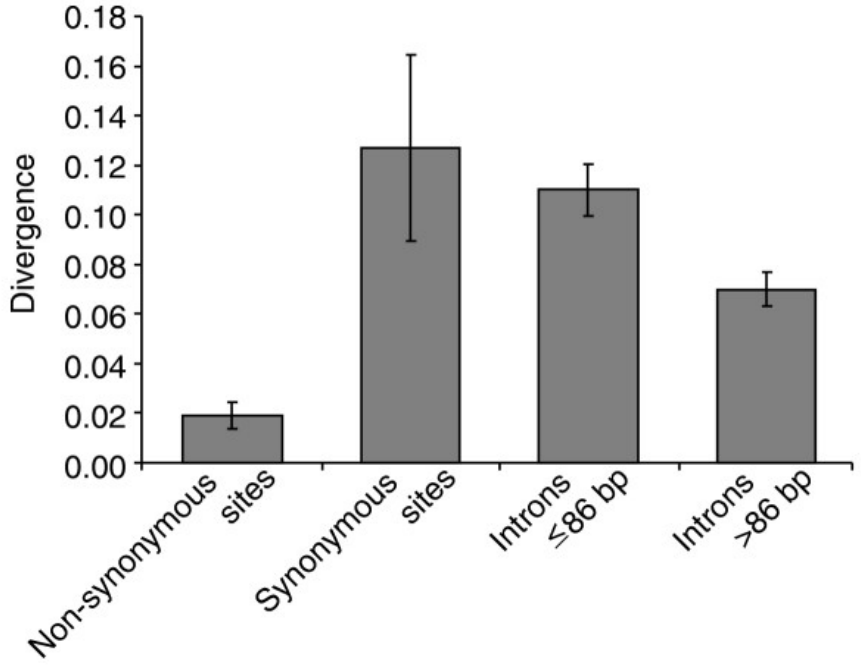


Evolutionary rates vary according to gene region

Drosophila

Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content

Penelope R Haddrill*, Brian Charlesworth*, Daniel L Halligan* and Peter Andolfatto†

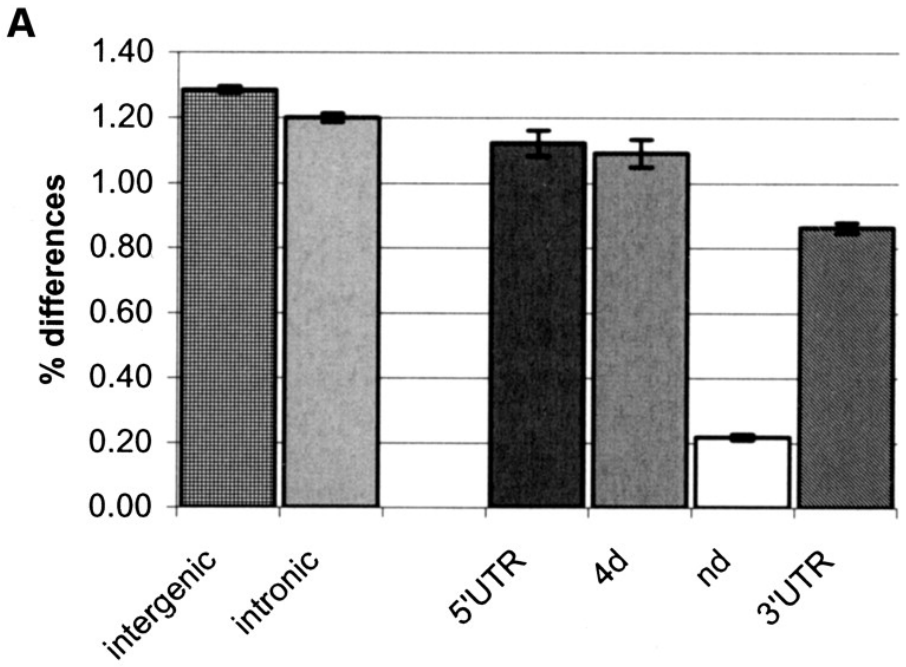


Human - Chimp

Selection on Human Genes as Revealed by Comparisons to Chimpanzee cDNA

Ines Hellmann, Sebastian Zöllner, Wolfgang Enard, et al.

Genome Res. 2003 13: 831-837
Access the most recent version at doi:10.1101/gr.944903



Molecular clock hypothesis

- JC and Kimura models assume nucleotides accrue substitutions at a constant rate
- Empirical evidence
- Useful concept for dating divergence times
- Deviation indicate slowing or acceleration of evolutionary change
- …or incorrect fossil dating

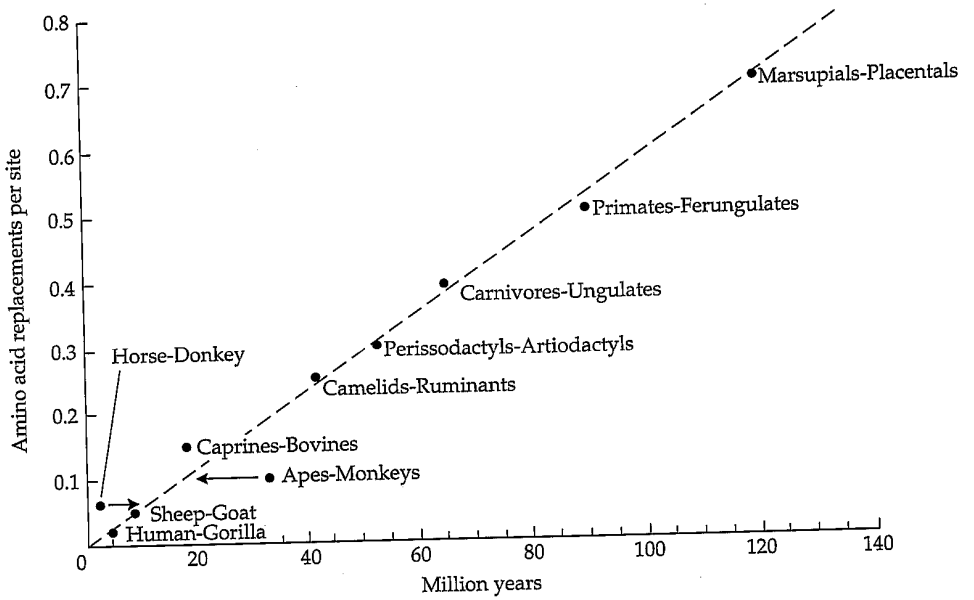


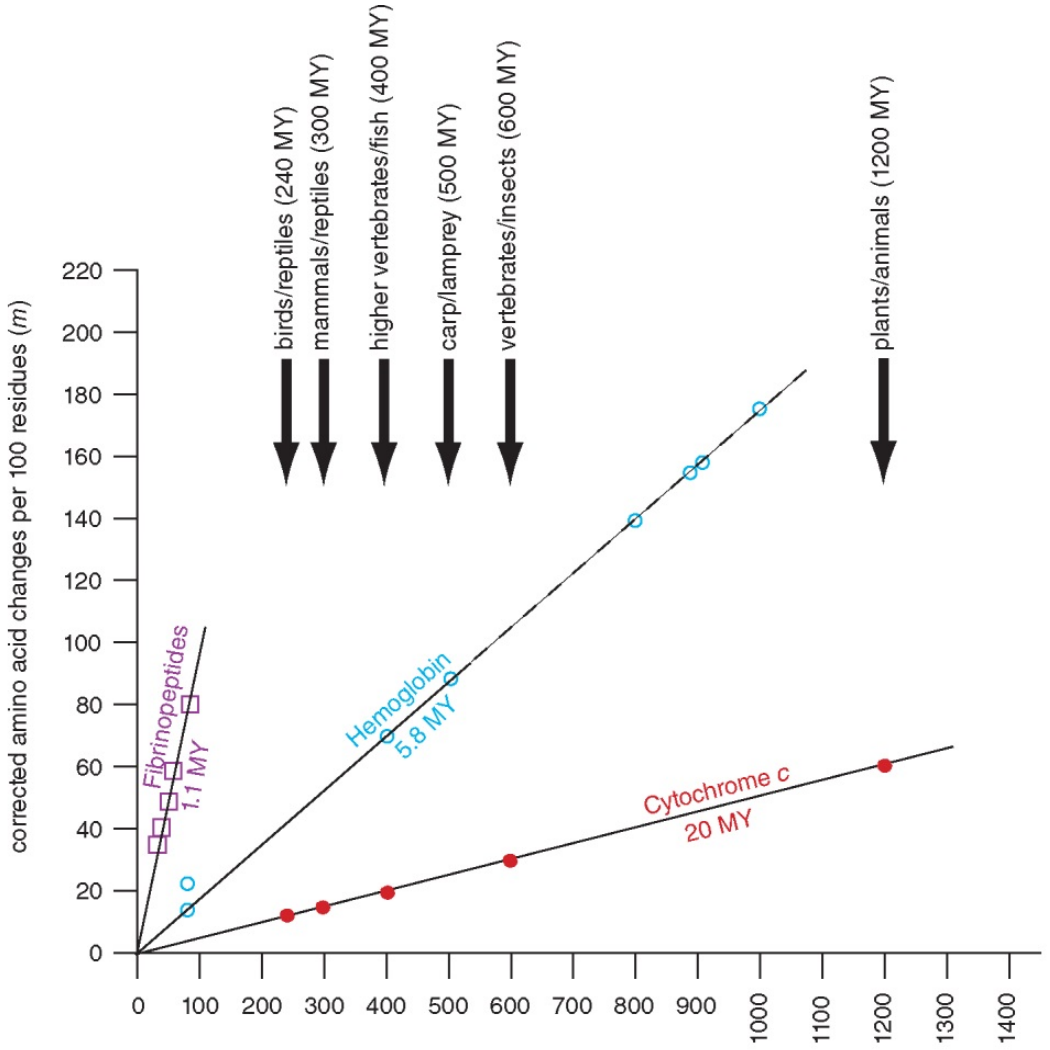
FIGURE 4.15 Number of amino acid replacements per amino acid site in a combined sequence consisting of hemoglobins α and β, cytochrome *c*, and fibrinopeptide A among various mammalian groups plotted against geological estimates of divergence times. The dashed line represents the molecular clock expectation of equal rates of amino acid replacement in all evolutionary lineages. There are two large deviations of the observed values from the expected line. These deviations indicate a slowdown in evolution following the divergence between apes and monkeys, and an acceleration following the divergence between horse and donkey. However, these inferences are based on specific paleontological estimates of divergence times (33 million years for the ape-monkey split and 2 million years for the horse-donkey split), and if these time estimates are inaccurate (arrows), the deviation of these lineages from a strict molecular clock may not be significant. Modified from Langley and Fitch (1974).

Molecular clock hypothesis

- Forces effecting sequence change
  - **Mutation**: sequence change in a single individual
  - **Fixation:** there exists at least two sequence variants (alleles) in a population and overtime only one remains
  - **Drift:** change in variant frequency due to resampling
  - **Selection**
    - Negative Selective removal of deleterious mutations (alleles)
    - Positive Increase the frequency of beneficial mutations (alleles) that increase fitness (success in reproduction)
- Main innovation: most changes we observe across lineages are neutral

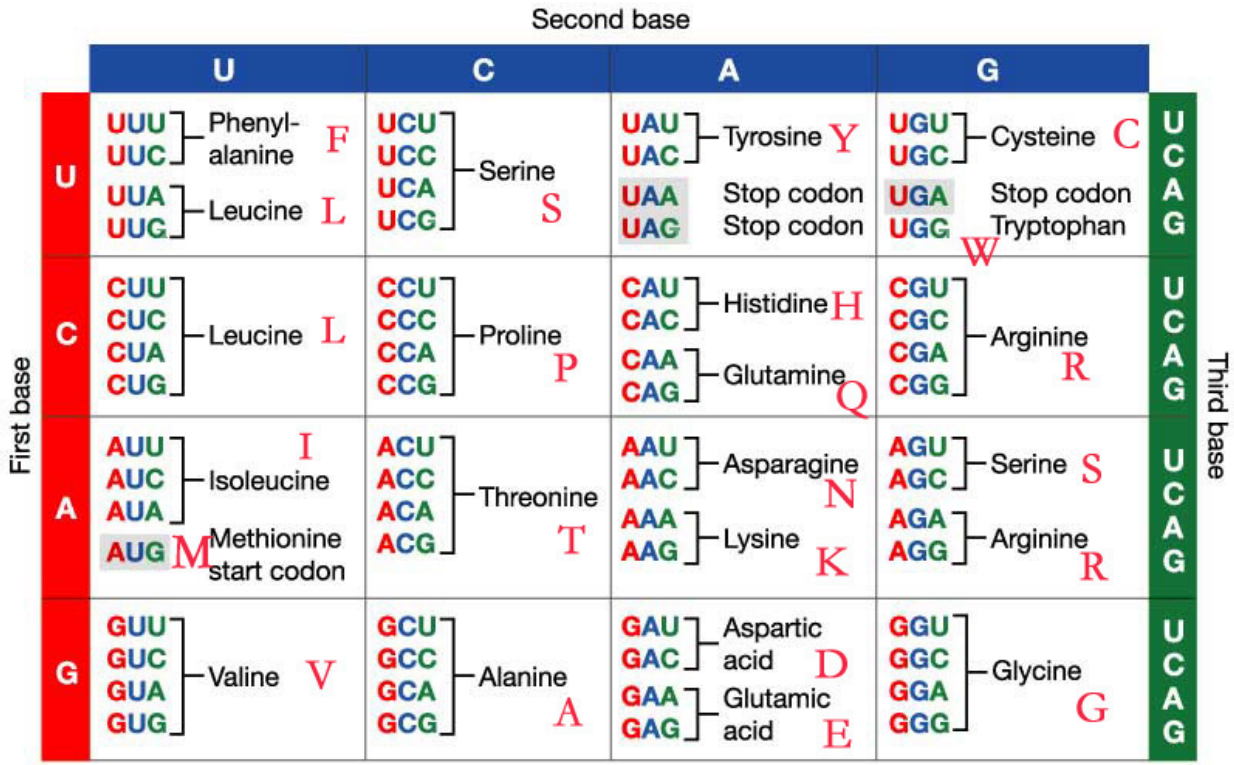
Molecular clock hypothesis

- Rates vary widely for different proteins but scale with time
- Local clock vs global clock
- Rates can vary over branches and over time
  - Selection
  - Generation time effect
  - Efficiency of DNA repair
    - Some evidence suggests that DNA repair is more efficient in humans than in mice



Protein-coding sequences present opportunities to study differential rates

- A **nonsynonymous** substitution is a nucleotide **mutation** that alters the amino acid sequence of a protein.
- **Synonymous** substitutions do not alter amino acid sequences.
- **Synonymous** (silent) changes are thought to have relatively small effects, if any, on gene and protein function
- **synonymous** sites typically diverge at rates similar to non-functional sequences, such as pseudogenes, they are often the best molecular clock to normalize rates of substitution.



Synonymous and nonsynonymous substitution rates K_S and K_A

Early methods estimated K_S and K_A using simple counting methods. These were sufficient as long as divergence was low (< 1 change per codon).

Step 1: count # syn and nonsyn changes (M_S & M_A)

Step 2: normalize each by number of syn and nonsyn sites (N_S & N_A)

- for each nucleotide, sum proportion of potential changes that are syn or nonsyn
- determine mean # syn and nonsyn sites between sequences

Step 3: use nucleotide model to compute genetic distances K_S and K_A

$K_S = -¾ \ln(1 – (4/3)*(M_S/N_S))$

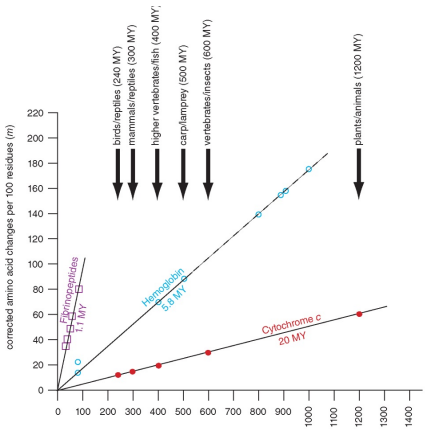
$K_A = -¾ \ln(1 – (4/3)*(M_A/N_A))$

More sophisticated models can be used to account for ts/tv rates.

Interpreting K_S and K_A

These quantities can be powerful for making inferences about protein function.

- While amino acid divergence rates between proteins vary 1,000 fold, it is not clear whether rapidly evolving regions resulted from a lack of functional constraint or from positive selection for novel function.
- We can distinguish between these two scenarios by normalizing K_A with K_S



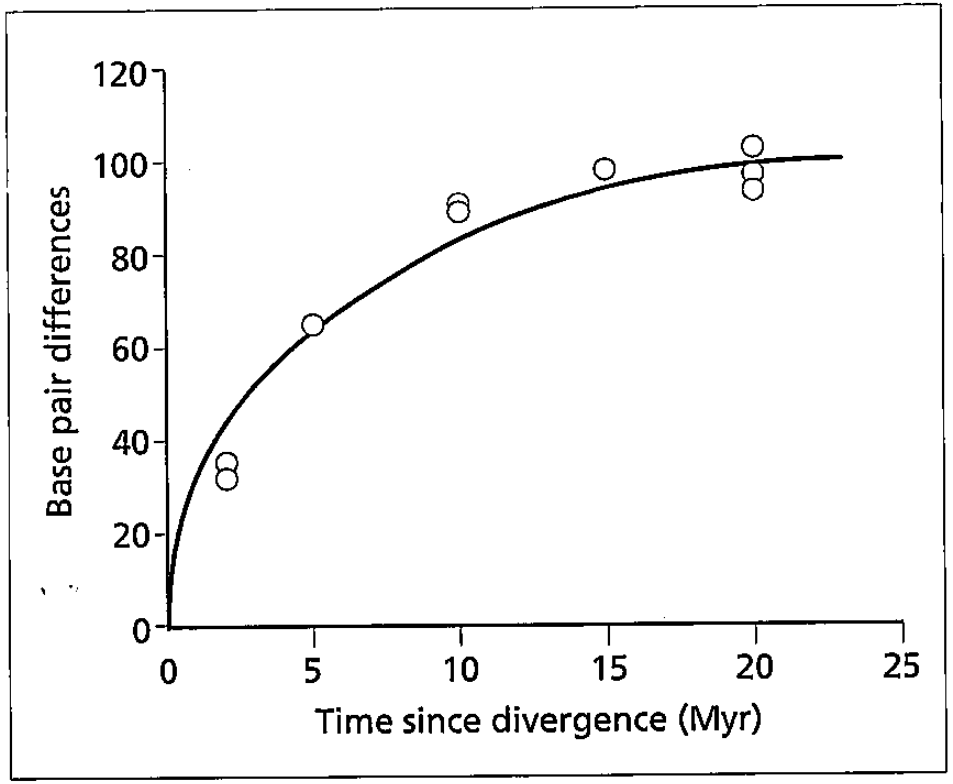
Statistical Tests

- $K_A - K_S$ use t-test to assess significance
  - $t = (K_A - K_S) / \text{sqrt}(V(K_A) + V(K_S))$, V is variance
  - Cannot be used for small numbers of substitutions $M<10$
- 2 x 2 Contingency table and Fisher's exact test
  - Because M_A and M_S are not corrected genetic distances
  - this is only accurate for small numbers of substitutions ~ $K < 0.2$
- K_A / K_S (a.k.a. d_N/d_S) It is convenient to compare them as a ratio, in which the nonsyn rate is normalized by the syn rate. These values are comparable across genes and species.

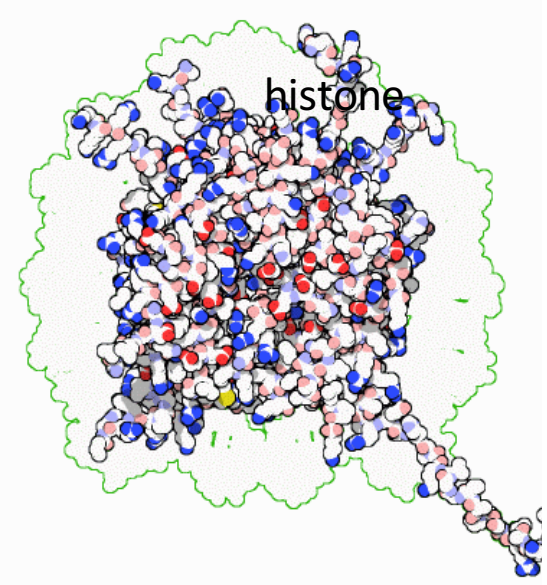
	nonsyn	syn
changed	M_A	M_S
not changed	N_A-M_A	N_S-M_S

dN/dS in practice

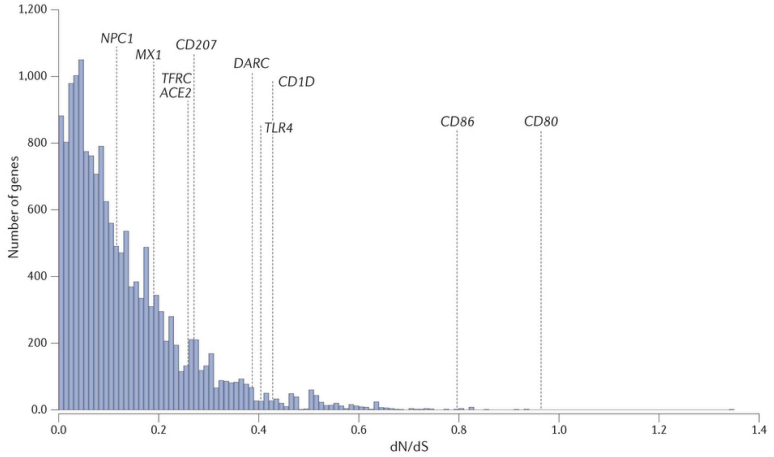
- Only useful for comparing close sequences
- Mammalian genes –YES!
- Yeast family genes –NO
- Synonymous sites are at saturation!



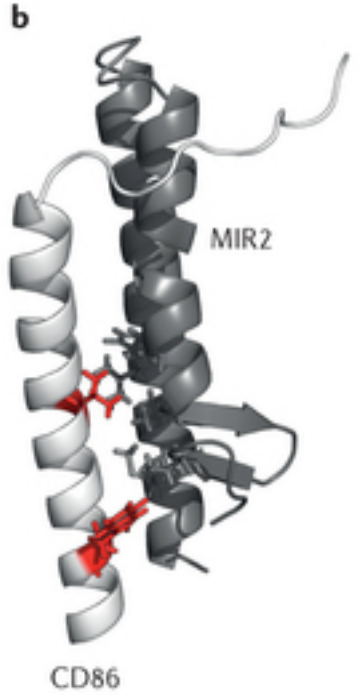
dN/dS genome wide



Nucleosomes
Cytochrome c oxidase
Rad51



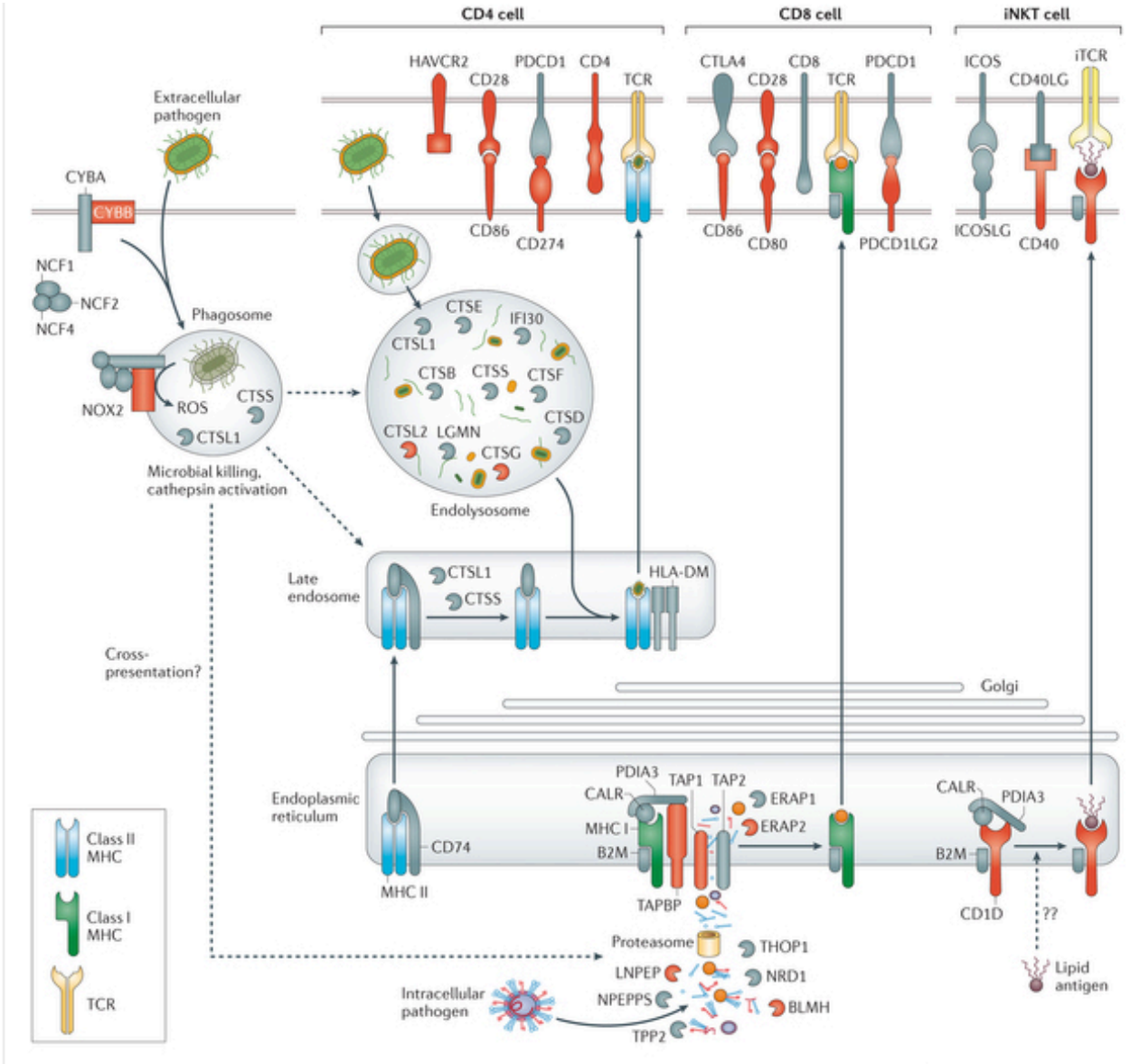
Pairwise comparisons between mouse and human genes.
(Evolutionary insights into host–pathogen interactions from mammalian sequence data)



Pathogen defense
Fertilization proteins
Olfactory receptors
Detoxification enzymes

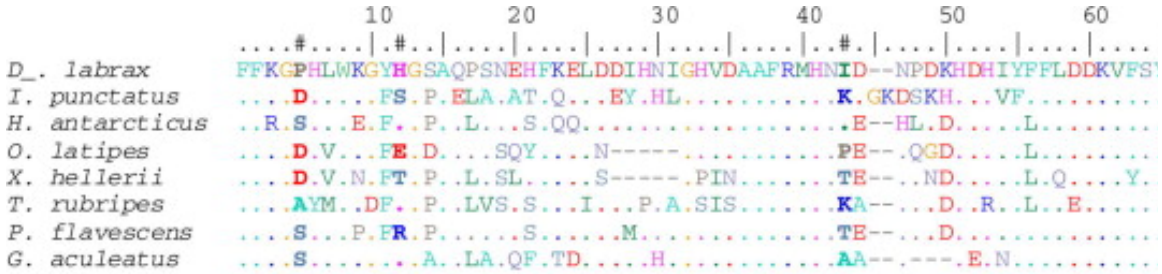
Only rarely do d_N/d_S ratios calculated over the entire gene exceed 1.
Usually only select protein regions experience recurrent positively selected changes, so that moderately elevated values may indicate the presence of positive selection.

Immune proteins are targets of positive selection

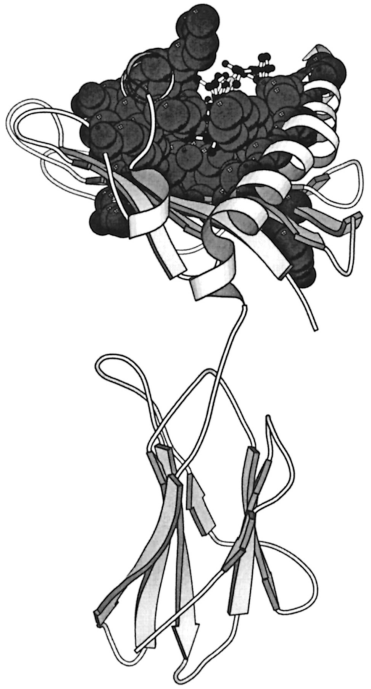


Example MHC

- MHC molecules bind intercellular peptides and present them to immune cells
- Important for recognizing virus infected cells
- Overall d_N/d_S ratio <1



Amino acids in **spacefill** have a d_N/d_S **ratio > 1**.



Reproductive proteins are often under positive selection

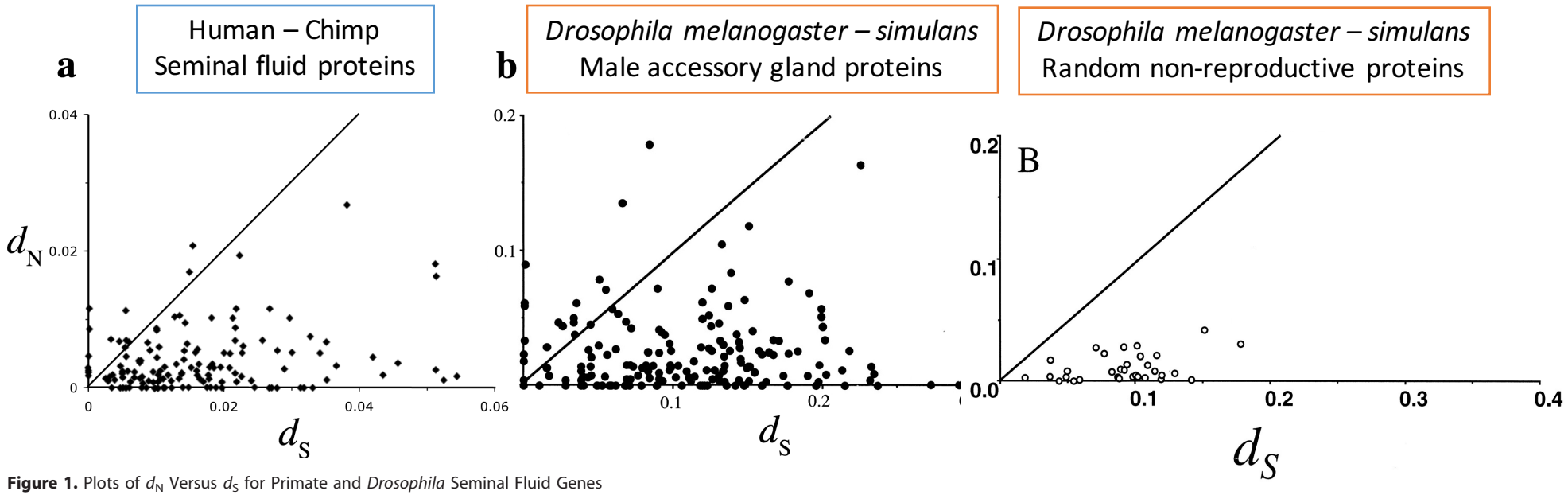


Figure 1. Plots of d_N Versus d_S for Primate and *Drosophila* Seminal Fluid Genes

(A) Genes encoding seminal fluid proteins identified by mass spectrometry in human versus chimpanzee.
(B) *Drosophila simulans* male-specific accessory gland genes versus *D. melanogaster* [2].
The diagonal represents neutral evolution, a d_N/d_S ratio of one. Most genes are subject to purifying selection and fall below the diagonal, while several genes fall above or near the line suggesting positive selection. Comparison of the two plots shows elevated d_N/d_S ratios in seminal fluid genes of both taxonomic groups.
DOI: 10.1371/journal.pgen.0010035.g001

Particular protein functional classes demonstrate frequent signs of positive selection

From Human-Macaque gene comparisons

S6.7: PANTHER Categories Showing an Excess of Positive Selection

Description	N^a	$P_A{}^b$
Defense/immunity protein	141	**1.99e-16**
Immunoglobulin receptor family member	44	**2.33e-11**
Immunity and defense	608	**2.59e-06**
Natural killer cell mediated immunity	28	**3.20e-06**
Fertilization	16	**2.64e-05**
KRAB box transcription factor	263	**3.83e-05**
Intermediate filament	41	**6.01e-05**
Other receptor	95	**1.70e-04**
Structural protein	88	**2.66e-04**
Complement-mediated immunity	27	4.80e-04
Complement component	23	9.65e-04
Blood clotting	35	3.18e-03

Example:lysozymes

Lysozyme is a bacteriolytic enzyme normally acting in host defense. It has been coopted to the foregut in vertebrate species that digest plant material – namely ruminants (e.g. cow), colobine monkeys (e.g. langur), and the hoatzin, a bird.

Lysozymes in these species have independently **converged** to the same amino acid at specific sites to the effect of increasing tolerance to the low pH of the digestive tract.

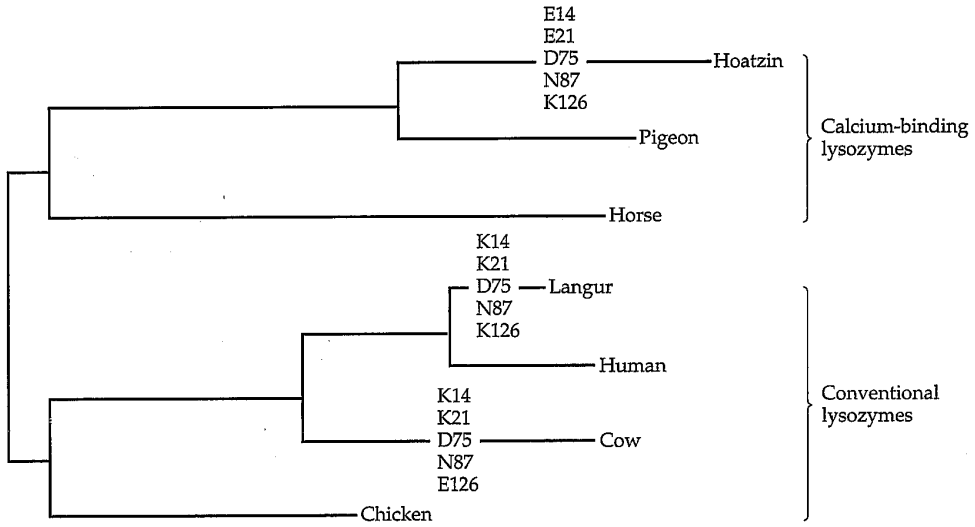


Table 2
Structural Adaptations Evident in Hoatzin Stomach Lysozyme

	Lysozyme Type			
Characteristic	Hoatzin Stomach	Mammalian Stomach	Chicken Egg-White	Pigeon Egg-White
Low pH optimum	+	+	−	−
Isoelectric point	~6	6.2–7.7[a]	11.2	~10.6
Total arginines	5	3–6	11	10
Arginine-to-lysine ratio	0.63	0.27–0.67	1.83	0.77
Adaptive residues:				
E/K14	+	+	−	−
E/K21	+	+	−	−
D75	+	+	−	−
N87	+	+	−	−
E/K126	+	+	−	−

Note.—Mammalian stomach lysozymes used in this comparison were from three true ruminants (cow, goat, and axis deer) and the langur monkey. Adaptive replacements are summarized in the single-letter amino acid code and numbered according to fig. 3. Properties of chicken and pigeon lysozymes are shown to represent nonstomach lysozymes.

[a] The isoelectric point for langur lysozyme is somewhat higher (Stewart et al. 1987; Stewart and Wilson 1987).