Molecular Evolution

3-Feb-2014

DEKM book

Notes from Drs. B. John and T. Benos

Mathematical Models in Biology, An Introduction, Allman and Rhodes, CUP 2004.

Completed Genomes

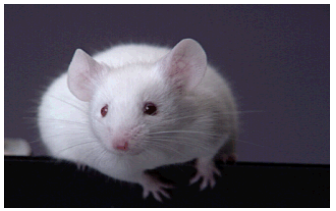




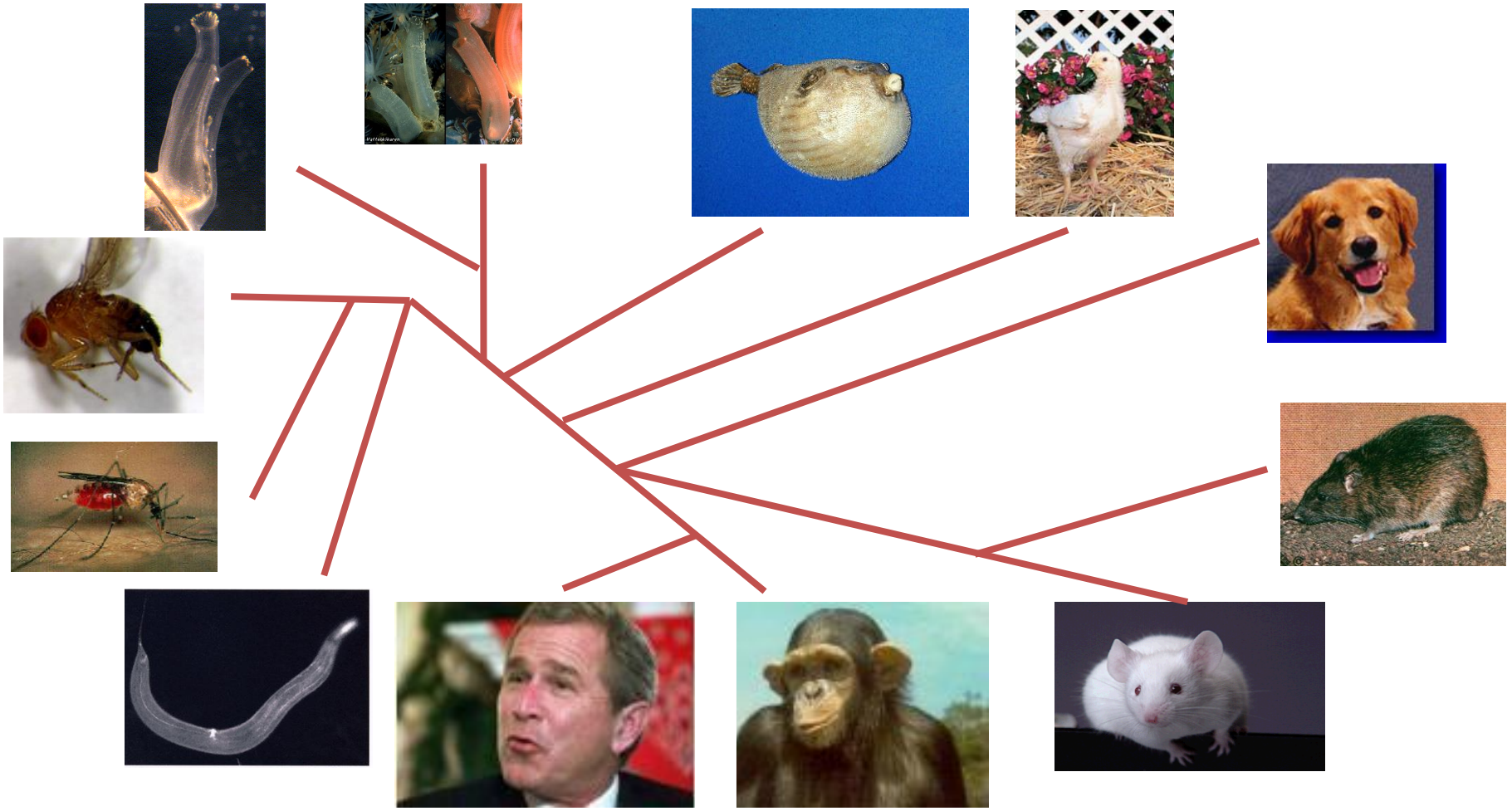

More than 200 complete genomes have been sequenced



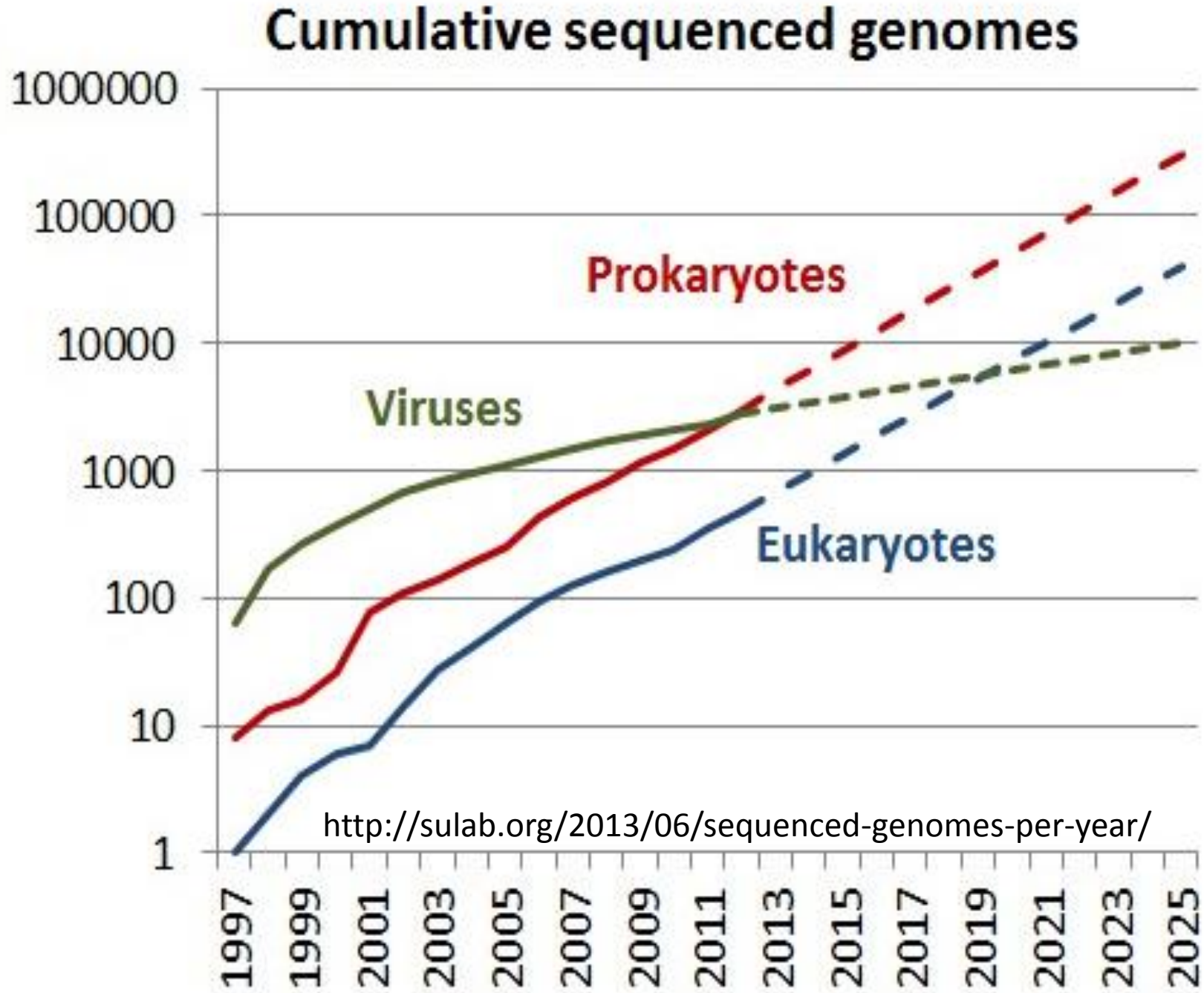


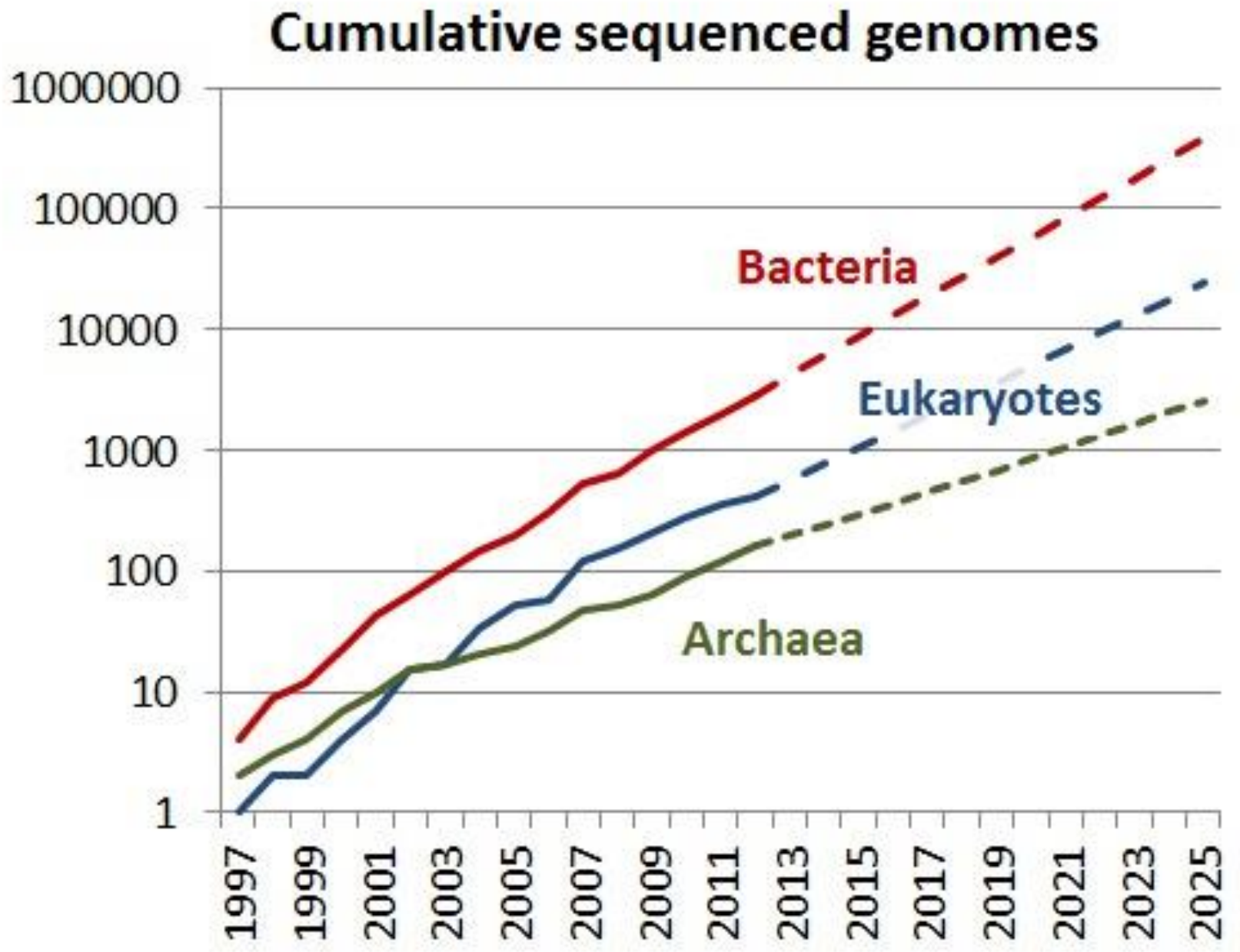
Evolution



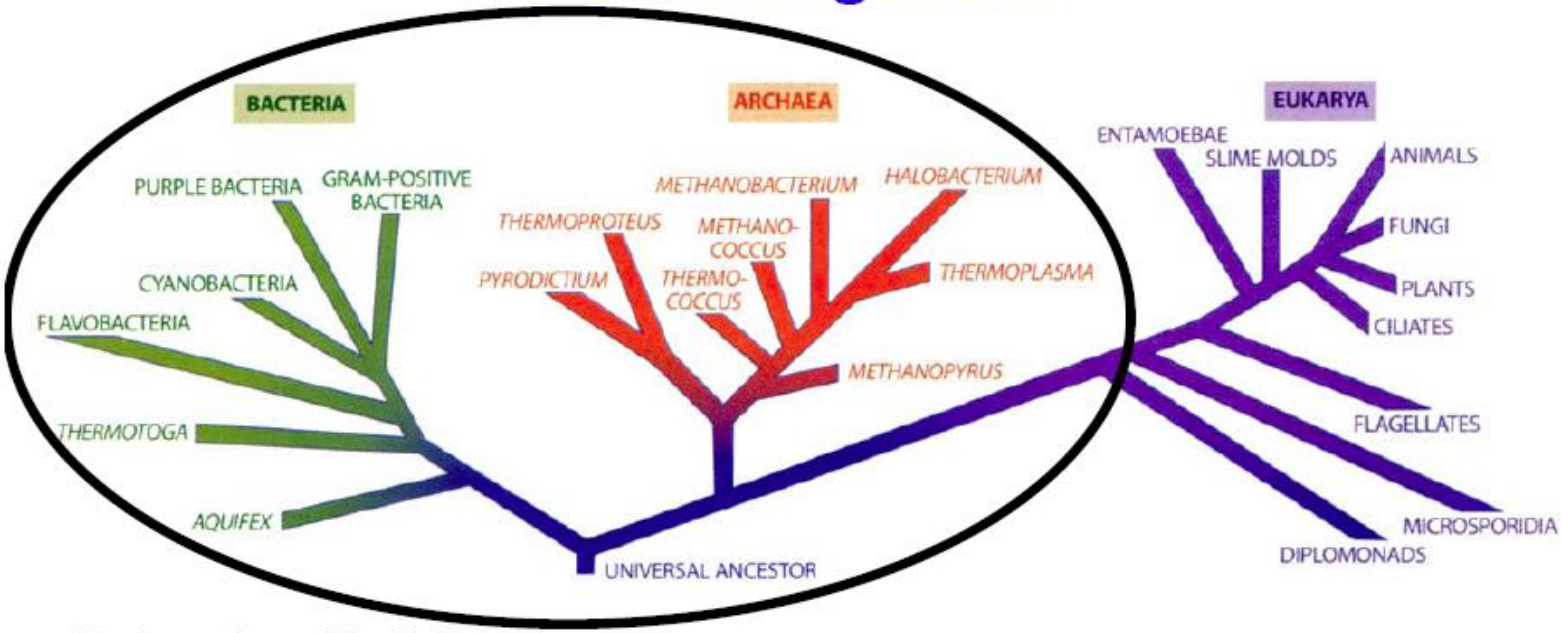
Slide courtesy: Serafim Batzoglou


Cumulative sequenced genomes
1000000
100000
10000
1000
100
10
1
Prokaryotes
Viruses
Eukaryotes
http://sulab.org/2013/06/sequenced-genomes-per-year/
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015
2017
2019
2021
2023
2025

http://sulab.org/2013/06/sequenced-genomes-per-year/



Three Kingdoms



Prokaryotes – No Cell nucleus,
No organelles

Eukaryotes – Cell nucleus

M. Madigan and B. Marrs, 1997

4000 Myr: life originated in the sea / shallow ponds
3500 Myr: bacteria emerged from the seas to colonise the land
1000 Myr: plants / fungi slowly colonized inland of coastal margins
700 Myr: insects / arthropods crawled out to feed on plants
350 Myr: vertebrates adapted to air breathing / survival on land

Viruses?

Evolutionary Taxonomy

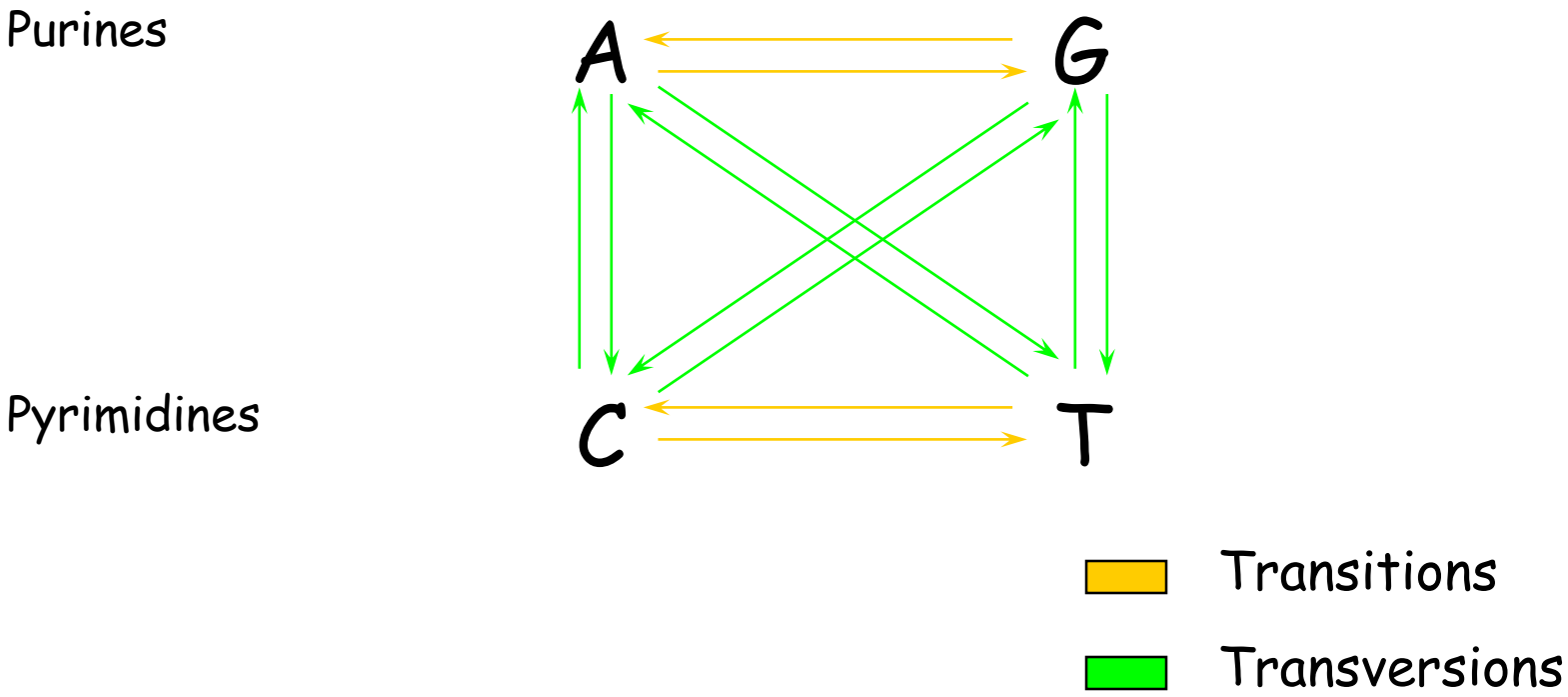
- Track DNA of species to determine phylogenetic relationship between species
- Assumptions
  - DNA mutates slightly over generations
  - Species descended from a common ancestor must have DNA sequences “similar” to each other

Examples

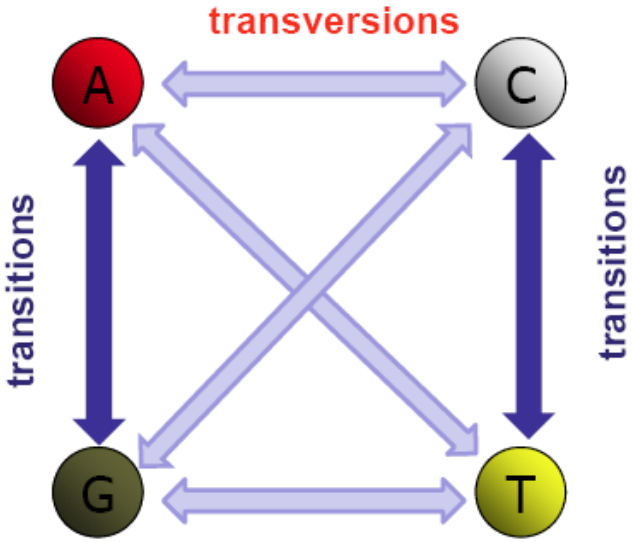
- Example sequence: CGTGACTTCC
  - Base substitution: CGGGACTTCC
  - Base insertion: CGATGACTTCC
  - Base deletion: CG GACTTCC
  - Sequence insertion: CGTATTAGGACTTCC
  - Sequence deletion: CG CTTCC
  - Sequence duplication: CGTGACTTGACTTCC
  - Sequence inversion: CGTTCAGTCC

Base mutations (general): definitions

- Base mutations: the source of sequence variation
- Transitions more frequent than transversions

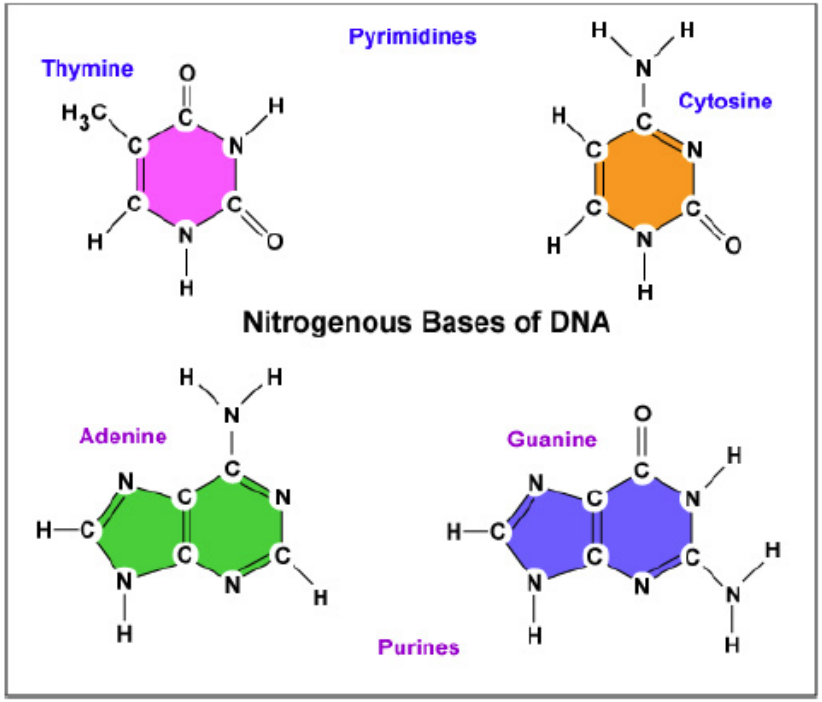


Molecular evolution via substitutions



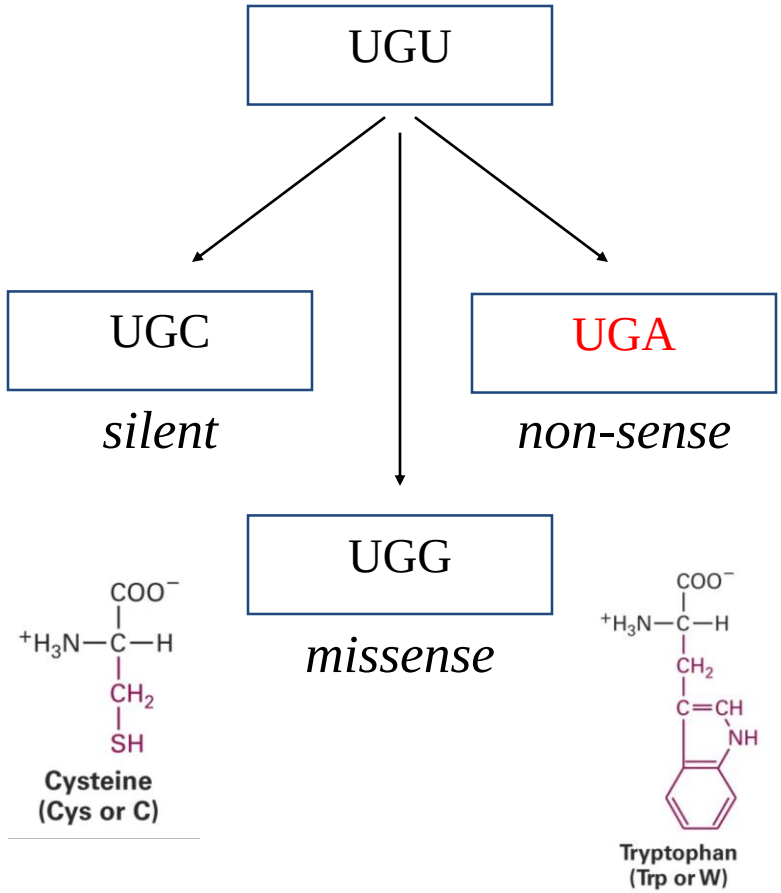
transitions– pyrimidine to pyrimidine

transversions – purine to pyrimidine or vice-versa



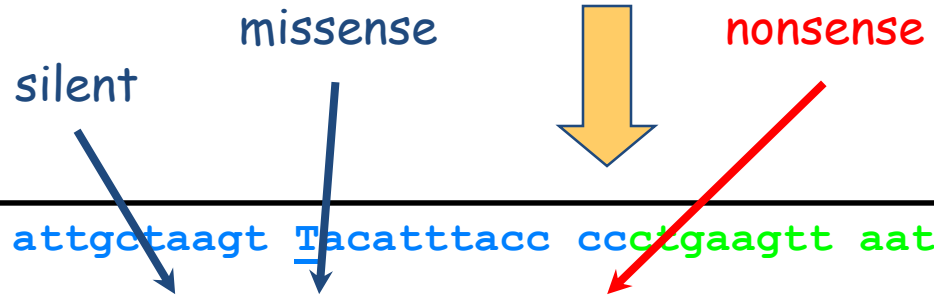
Base mutations in ORFs: definitions

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU UUC Phenylalanine	UCU UCC UCA UCG Serine	UAU UAC Tyrosine	UGU UGC Cysteine	U C A G
	UUA UUG Leucine		UAA UAG Stop codon Stop codon	UGA Stop codon; UGG Tryptophan	
C	CUU CUC CUA CUG Leucine	CCU CCC CCA CCG Proline	CAU CAC Histidine	CGU CGC CGA CGG Arginine	U C A G
			CAA CAG Glutamine		
A	AUU AUC AUA Isoleucine	ACU ACC ACA ACG Threonine	AAU AAC Asparagine	AGU AGC Serine	U C A G
	AUG Methionine; initiation codon		AAA AAG Lysine	AGA AGG Arginine	
G	GUU GUC GUA GUG Valine	GCU GCC GCA GCG Alanine	GAU GAC Aspartic acid	GGU GGC GGA GGG Glycine	U C A G
			GAA GAG Glutamic acid		



Base mutations in ORFs: definitions (cntd)

```
tggagctAtt attgctaagt Aacatttacc ccctgaagtt aatgGatcaa tcaagagaga 120
tgtgggctgt aatgaaTcgt CttattgaaT TaacaggttG gatcgttctt gtcgtttcag 180
            M   N   R   L   I   E   L
tcattcttct tggcgtggcg agtcacattg acaactatca gccacctgaa cagagtgctt 240
cggtacaaca caagtaagct ctgcacttgt ggagcgacat gctgcccgtc cgggtgcatg 300
```




```
tggagctGtt attgctaagt Tacatttacc ccctgaagtt aatgAatcaa tcaagagaga 120
tgtgggctgt aatgaaCcgt GttattgaaT Aaacaggttg gatcgttctt gtcgtttcag 180
            M   N   R   V   I   E
tcattcttct tggcgtggcg agtcacattg acaactatca gccacctgaa cagagtgctt 240
cggtacaaca caagtaagct ctgcacttgt ggagcgacat gctgcccgtc cgggtgcatg 300
```

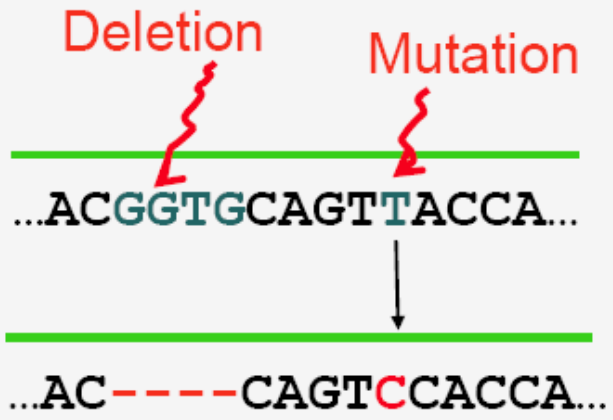
Base mutations in ORFs: definitions (cntd)

```
tggagctAtt attgctaagt Aacatttacc ccctgaagtt aatgGatcaa tcaagagaga 120
tgtgggctgt aatgaaTcgt CttattgaaT Taacaggttg gatcgttctt gtcgtttcag 180
              M  N  R  L  I  E  L
tcattcttct tggcgtggcg agtcacattg acaactatca gccacctgaa cagagtgctt 240
cggtacaaca caagtaagct ctgcacttgt ggagcgacat gctgcccgtc cgggtgcatg 300
```

deletion

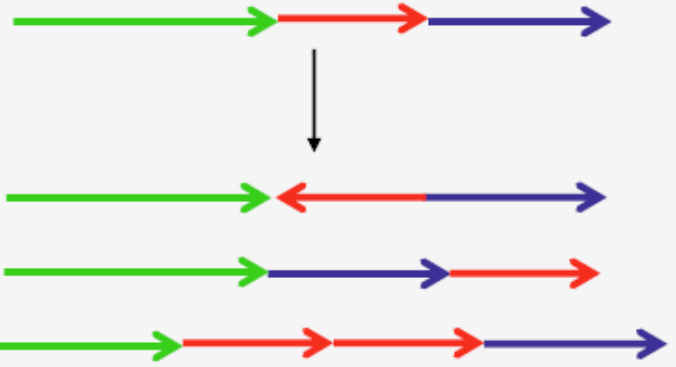
```
tggagctGtt attgctaagt Tacatttacc ccctgaagtt aatgAatcaa tcaagagaga 120
tgtgggctgt aatgaaCcgt Gttattgaa- Taacaggttg gatcgttctt gtcgtttcag 180
              M  N  R  V  I  E
tcattcttct tggcgtggcg agtcacattg acaactatca gccacctgaa cagagtgctt 240
cggtacaaca caagtaagct ctgcacttgt ggagcgacat gctgcccgtc cgggtgcatg 300
```

Evolution at the DNA level



SEQUENCE EDITS

REARRANGEMENTS



Molecular Evolution and its consequences

1. Most related sequences have many positions that have mutated several times
2. Rate of accepted mutation is usually not the same for all types of base substitution
3. Different codon position have different mutation rates
4. Influence of selective pressure on the observed frequency of synonymous and non-synonymous mutations

Most related sequences have many positions that have mutated several times

- Neither of the following is true
  - Sequences have only diverged to a moderate degree such that no position has been subjected to more than one mutation
    - If so, once the sequences are aligned, all mutational events could be observed as nonidentical aligned bases and assume mutation to be from one base to another
  - All sequences evolved at a constant mutation rate for all mutations at all times
    - If so, the number of observed differences between any two aligned sequences would be directly proportional to the time elapsed since they diverged from their most recent common ancestor
- Evolutionary distance = *p*-distance (fraction of misaligned residues)
  - Because of overlapping mutations, p-distance is an underestimate of the number of mutations that actually occurred

How many mutations?

- Knowing DNA seq info, measure the mutation amount in the evolutionary descent
- ancestor seq S0: ACCTGCGCTA
- intermediate seq S1: ACGTGCACTA
- descendent seq S2: ACGTGCGCTA
- (S0,S2) => 1 mutation => 1/10
- (S0,S1,S2) => 3 mutations => 3/10
- G -> A -> G is a “hidden” mutation

Rate of accepted mutation is usually not the same for all types of base substitution

- Simplest model: rates identical and time-invariant with no substitution preferences
- Whether a mutation is retained or lost from populations' gene pool will depend on many factors including..
  - AA sequence is altered or not
  - Effect on function

  $\Rightarrow$Rate of mutation and substitution preferences can vary at each position along the genome

R = transitions/tranversions

- R = ½ but 4 in practice (mitochondrial gene sequences from mammalian subfam Bovinae)

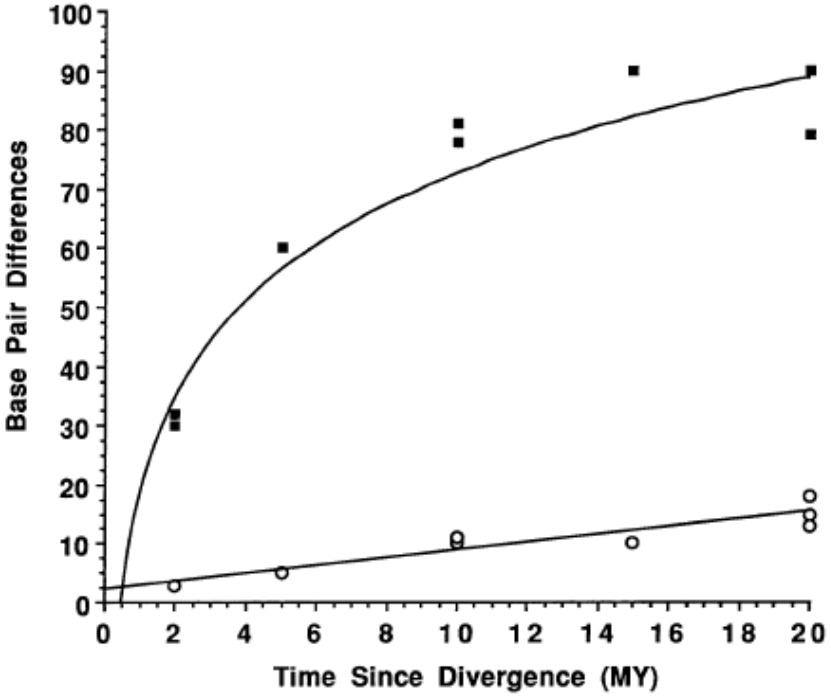
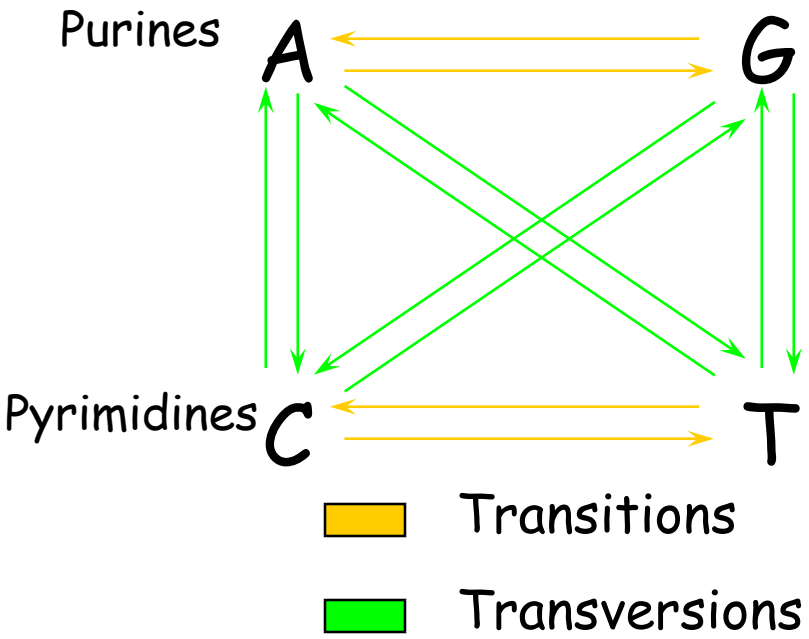


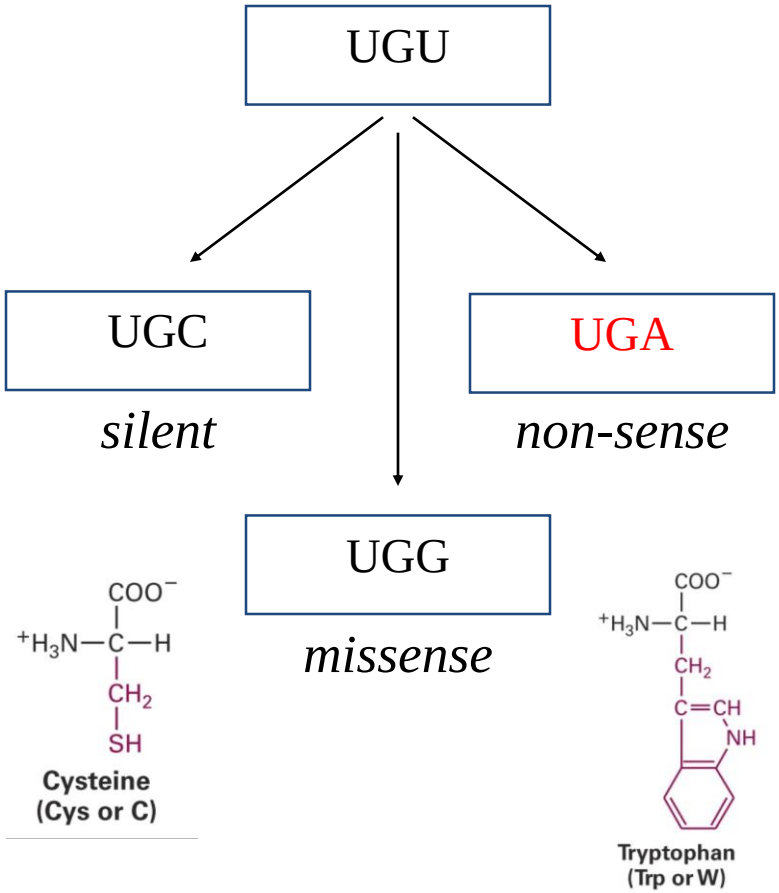
FIG. 3. Total nucleotide substitutions (squares) vs transversion differences (circles) between pairs of bovid taxa plotted against estimates of time since divergence. Taxon pairs and references for divergence times are listed in Table 2.



Janacek et al Mol. Phylogenet. Evol. 6: 107-119, 1996

Different codon positions have different mutation rates

First letter	Second letter: U	C	A	G	Second letter
U	UUU UUC Phenyl-alanine; UUA UUG Leucine	UCU UCC UCA UCG Serine	UAU UAC Tyrosine; UAA UAG Stop codon Stop codon	UGU UGC Cysteine; UGA Stop codon; UGG Tryptophan	U C A G
C	CUU CUC CUA CUG Leucine	CCU CCC CCA CCG Proline	CAU CAC Histidine; CAA CAG Glutamine	CGU CGC CGA CGG Arginine	U C A G
A	AUU AUC AUA Isoleucine; AUG Methionine; initiation codon	ACU ACC ACA ACG Threonine	AAU AAC Asparagine; AAA AAG Lysine	AGU AGC Serine; AGA AGG Arginine	U C A G
G	GUU GUC GUA GUG Valine	GCU GCC GCA GCG Alanine	GAU GAC Aspartic acid; GAA GAG Glutamic acid	GGU GGC GGA GGG Glycine	U C A G



Synonymous mutations almost always from third codon position => mutation rate will be higher

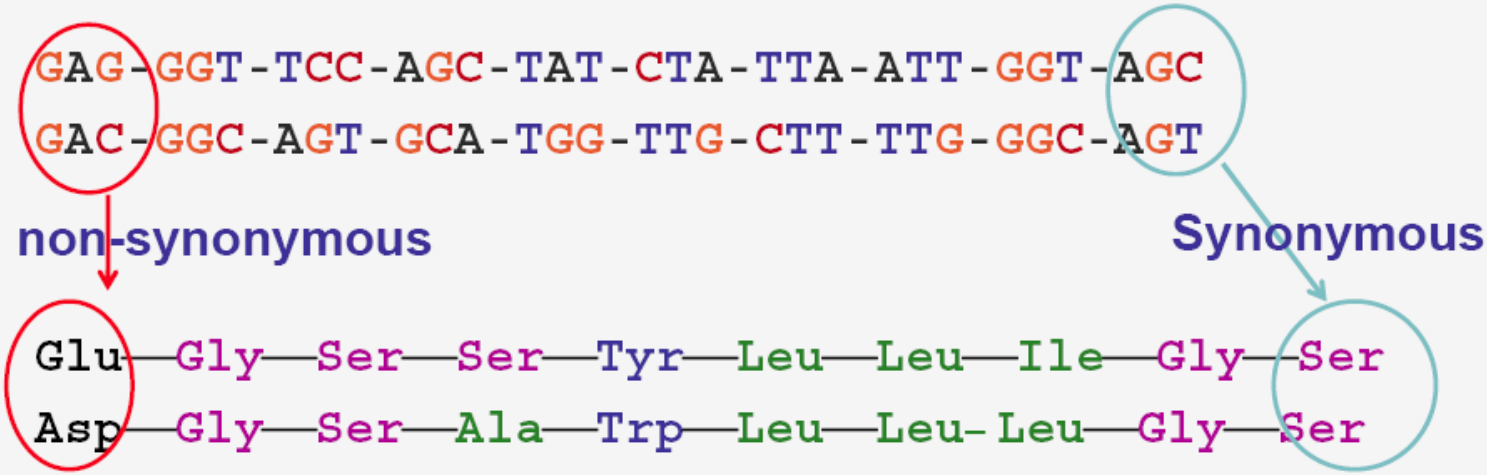
Influence of selective pressure on the observed frequency of synonymous and non-synonymous mutations

- Mutation: maintain or loose
- Depends on the selective pressure on the species and on whether the fitness of the organism changes because of mutation
- No selective pressure => random genetic drift
- Fitness: positive selection (mutation kept) vs negative selection (mutation lost)

Influence of selective pressure on the observed frequency of synonymous and non-synonymous mutations

- With aligned sequences, possible to identify the type of selection
  - Synonymous mutations will not affect fitness, thus neither selected for nor against…=> null model
  - Ratio of the rates of non-synonymous mutations/synonymous mutations to determine positive, negative or neutral selection

Evolution – Synonymous vs non-synonymous mutations for protein-coding regions



D_N= # of Non-synonymous mutations
D_S = # of Synonymous mutations

d_N= rate of non-synonymous mutations
d_S= rate of synonymous mutations

To get the rates, we need to apply correction! We will talk more on this shortly.

Tests for selection on sequences (rule of thumb)

$$\frac{d_N}{d_S} = 1 \text{ when replacement substitions are neutral}$$

$$\frac{d_N}{d_S} < 1 \text{ when replacement substitions are deleterious}$$

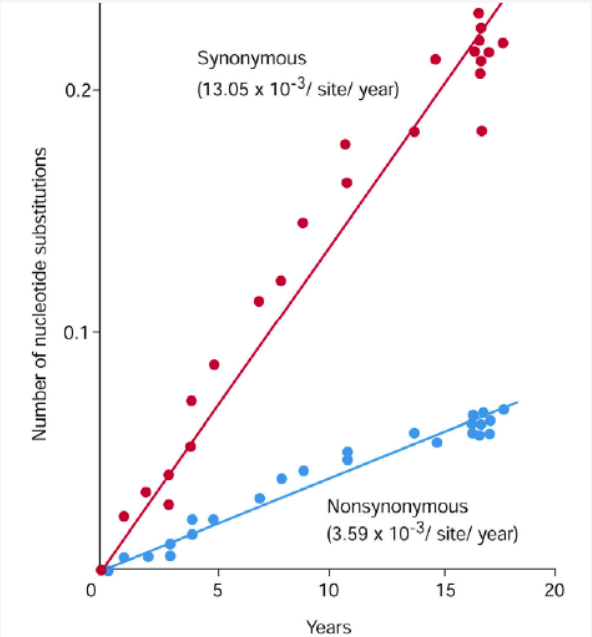
$$\frac{d_N}{d_S} > 1 \text{ when replacement substitions are advantageous}$$

Also denoted as Ka/Ks test

An example of Neutral theory

Prediction:

Synonymous substitutions accumulate in genomes faster than do non-synonymous ones



Influenza virus evolution over 20 years

How many mutations?

- Knowing DNA seq info, measure the mutation amount in the evolutionary descent
- ancestor seq S0: ACCTGCGCTA
- intermediate seq S1: ACGTGCACTA
- descendent seq S2: ACGTGCGCTA
- (S0,S2) => 1 mutation => 1/10
- (S0,S1,S2) => 3 mutations => 3/10
- G -> A -> G is a “hidden” mutation

Modeling mutations

- Assume mutations are very rare, so that no more than 1 mutations occurs at the same site..very restrictive
- Or use a suitable mathematical tool…probabilities
- Central question: Relate the observed fraction of sites that have mutated to the actual number of mutations that have occurred, which is not measurable from the data!!

Probability that next base is A?

- AGCCTACTGGCCAGGACCTC..
- Prob(A at site 21) = ?

Recovering hidden mutations

- Given a sequence, we are interested in Prob(purine|i) or Prob(pyrimidine|i)
- Assume in each generation, base i has 1.5% change of transversion =>
  - P(change|i) = 0.015 and P(no-change|i) = 0.985
- Assume changes within generations are independent to each other
- Consider changes over 2 generations with 4 possibilities: change/no-change, followed by change/no-change

Recovering hidden mutations

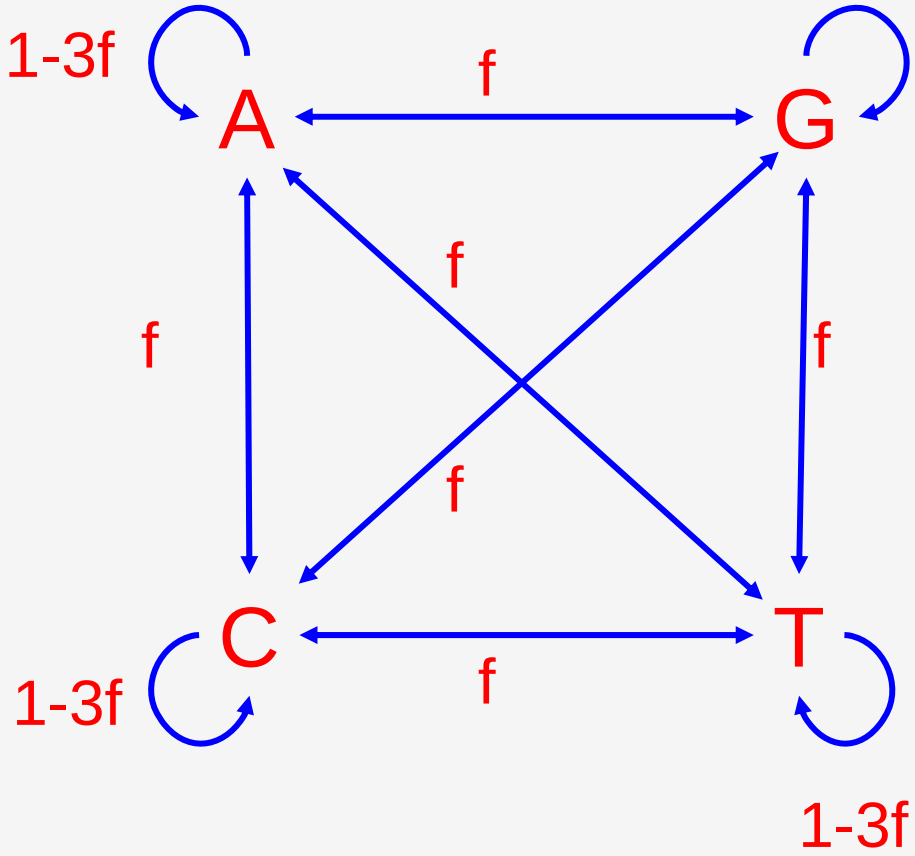
- P(change,change|i) = 0.015^2 = 0.000225
- P(change,no-change|i) =0.015 * 0.985 = 0.0148
- P(no-change,change|i) =0.985*0.015 = 0.0148
- P(no-change,no-change|i) = 0.985^2 = 0.970225
- Probability of hidden mutation: P(no-change,no-change|i) + P(change,change|i) = 0.970225 + 0.000225 = 0.97045

Mutations with Poisson model and distance correction

Central question: Relate the observed fraction of sites that have mutated to the actual number of mutations that have occurred, which is not measurable from the data!!

(blackboard)

Jukes Cantor model for estimating distances between sequences – Markov



$$A = \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} 1-3f & f & f & f \\ f & 1-3f & f & f \\ f & f & 1-3f & f \\ f & f & f & 1-3f \end{pmatrix} \end{array} \begin{array}{c} A \\ C \\ G \\ T \end{array}$$

Markov Transition probability Matrix

For markov models

pr(state after next is S_k | current state is S_i)

$= \sum_j pr$(state after next is S_k, next state is S_j | current state is S_i)

$= \sum_j pr$(next state is S_j| current state is S_i) x pr(state after next is S_k | current state is S_i, next state is S_j)

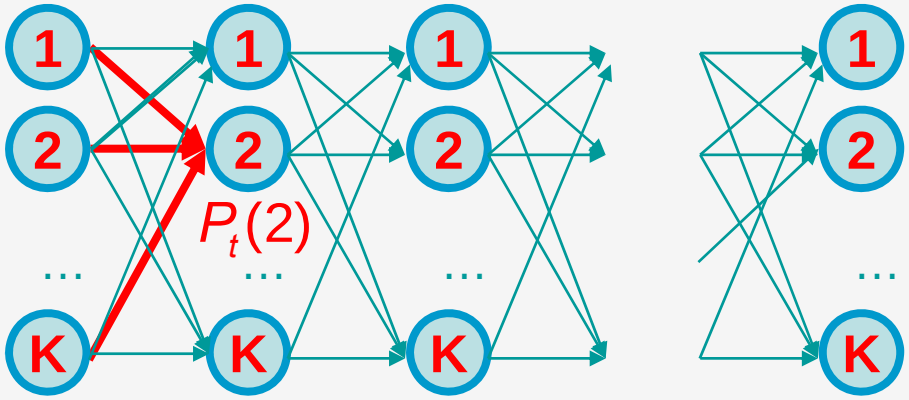
$= \sum_j p_{i,j} \; x \; p_{j,k}$

= (i,k)-element of P^2, where $P=(p_{i,j})$.

More generally,

pr(state t steps from now is S_k | current state is S_i) = i,k element of P^t

For Markov models



$$M_{ij}(t) = P_t(SP^t = j \mid SP^0 = i) = \sum_{k=1..K} P_t(SP^{t-1} = k, SP^t = j \mid SP^0 = i) =$$

$$\sum_{k=1..K} P_t(SP^{t-1} = k \mid SP^0 = i).P_t(SP^t = j \mid SP^{t-1} = k, SP^0 = i)$$

$$= \sum_{k=1..K} P_t(SP^{t-1} = k \mid SP^0 = i).P_t(SP^t = j \mid SP^{t-1} = k) =$$

$$\sum_{k=1..K} P_t(SP^{t-1} = k \mid SP^0 = i).P_t(SP^t = j \mid SP^{t-1} = k) = \sum_{k=1..K} M_{ik}(t-1).a_{kj} \Rightarrow$$

$$M(t) = M(t-1).A = M(t-2).A.A... \Rightarrow$$

$M(t) = \mathrm{M}(0).A^t = A^t$; since M(0)=Identity Matrix in DNA/protein evolution case

$$\Rightarrow M(t+d) = A^t.A^d = M(t).M(d)$$

Jukes-Cantor model - Formulation

$$M(t) = \begin{array}{c} \begin{matrix} \text{A} & \text{C} & \text{G} & \text{T} \end{matrix} \\ \begin{pmatrix} 1-3f(t) & f(t) & f(t) & f(t) \\ f(t) & 1-3f(t) & f(t) & f(t) \\ f(t) & f(t) & 1-3f(t) & f(t) \\ f(t) & f(t) & f(t) & 1-3f(t) \end{pmatrix} \end{array} \begin{matrix} A \\ C \\ G \\ T \end{matrix}$$

$$M(t+\Delta t) = M(t).M(\Delta t)$$

$$\Rightarrow M'(t) = \lim_{\Delta t \to 0} \frac{M(t+\Delta t) - M(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{M(t).M(\Delta t) - M(t).M(0)}{\Delta t} = M(t).M'(0)$$

Jukes-Cantor model - Formulation

$$M(t) = \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} 1-3f(t) & f(t) & f(t) & f(t) \\ f(t) & 1-3f(t) & f(t) & f(t) \\ f(t) & f(t) & 1-3f(t) & f(t) \\ f(t) & f(t) & f(t) & 1-3f(t) \end{pmatrix} \end{array} \begin{array}{c} A \\ C \\ G \\ T \end{array}$$

t^{th} step probability distribution matrix

OR

t-step transition matrix because (I,J) element is the Prob. To move from state I to J in t-steps

$$M(t+\Delta t) = M(t).M(\Delta t)$$

$$\Rightarrow M'(t) = \lim_{\Delta t \to 0} \frac{M(t+\Delta t) - M(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{M(t).M(\Delta t) - M(t).M(0)}{\Delta t} = M(t).M'(0)$$

JC model derivation- cont'd

$$M'(t) = M(t).M'(0) =$$

$$\begin{pmatrix} 1-3f(t) & f(t) & f(t) & f(t) \\ f(t) & 1-3f(t) & f(t) & f(t) \\ f(t) & f(t) & 1-3f(t) & f(t) \\ f(t) & f(t) & f(t) & 1-3f(t) \end{pmatrix} \begin{pmatrix} -3f'(0) & f'(0) & f'(0) & f'(0) \\ f'(0) & -3f'(0) & f'(0) & f'(0) \\ f'(0) & f'(0) & -3f'(0) & f'(0) \\ f'(0) & f'(0) & f'(0) & -3f'(0) \end{pmatrix}$$

$$\Rightarrow f'(t) = f'(0) - 3f(t)f'(0) - 3f(t)f'(0) + f(t)f'(0) + f(t)f'(0)$$

$$= f'(0) - 4f(t)f'(0) = \alpha - 4\alpha f$$

$$\Rightarrow f' = \alpha - 4\alpha f$$

$$\Rightarrow \int \frac{df}{\alpha - 4\alpha f} = \int dt = t$$

$$\Rightarrow f = \frac{1 - e^{-4\alpha t}}{4}$$ by assigning f(0) to 0

JC model derivation- cont'd

$$M'(t) = M(t).M'(0) =$$

$$\begin{pmatrix} 1-3f(t) & f(t) & f(t) & f(t) \\ f(t) & 1-3f(t) & f(t) & f(t) \\ f(t) & f(t) & 1-3f(t) & f(t) \\ f(t) & f(t) & f(t) & 1-3f(t) \end{pmatrix} \begin{pmatrix} -3f'(0) & f'(0) & f'(0) & f'(0) \\ f'(0) & -3f'(0) & f'(0) & f'(0) \\ f'(0) & f'(0) & -3f'(0) & f'(0) \\ f'(0) & f'(0) & f'(0) & -3f'(0) \end{pmatrix}$$

$$\Rightarrow f'(t) = f'(0) - 3f(t)f'(0) - 3f(t)f'(0) + f(t)f'(0) + f(t)f'(0)$$

$$= f'(0) - 4f(t)f'(0) = \alpha - 4\alpha f$$

$$\Rightarrow f' = \alpha - 4\alpha f$$

$$\Rightarrow \int \frac{df}{\alpha - 4\alpha f} = \int dt = t$$

$$\Rightarrow f = \frac{1 - e^{-4\alpha t}}{4}$$ by assigning f(0) to 0

JC model – cont'd

$$f = \frac{1-e^{-4\alpha t}}{4} = P_{ij}, i \neq j$$

$$and\ 1\text{-}3f = \frac{1+3e^{-4\alpha t}}{4} = P_{ij}, i = j$$

the probability that two sequences differ at a given position,

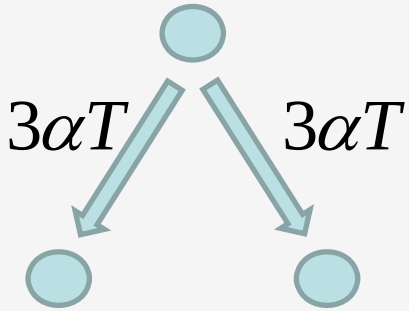
=Total mismatches/sequence length or D

=1- prob. that the positions are identical

1- prob. that the positions (stay the same + both change to the same nt)

$$= 1-\{P_{AA}^2 + P_{AT}^2 + P_{AG}^2 + P_{AC}^2\} = \frac{3}{4}(1-e^{-8\alpha t})$$

JC model – last step



Total number of substitutions per site=2x3αt=6αt = K

$$Since\ D = \frac{3}{4}(1 - e^{-8\alpha t}) \Rightarrow D = \frac{3}{4}(1 - e^{-4K/3})$$

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

$K.n = corrected\ \ distance$

Illustration

Human and bovine beta-globins are aligned with no deletions at 145 out of 147 sites. They differ at 23 of these sites. Thus $n_{\neq}/n = 23/145$, and the corrected distance using the Jukes-Cantor formula is (natural logs)

$$-19/20 \times \log(1 - 20/19 \times 23/145) = 17.3 \times 10^{-2} \text{ per site}$$

$$\Rightarrow 17.3 \times 10^{-2} \times 145 = 25.1$$

$$K = \frac{-19}{20} \ln(1 - \frac{20}{19} d);$$

$$K * 100$$

Corrected distances between protein sequences

- Below diagonal: observed number of differences per 100 amino acids
- Above diagonal: number of differences per 100 amino acids
- Correction method: Jukes-Cantor

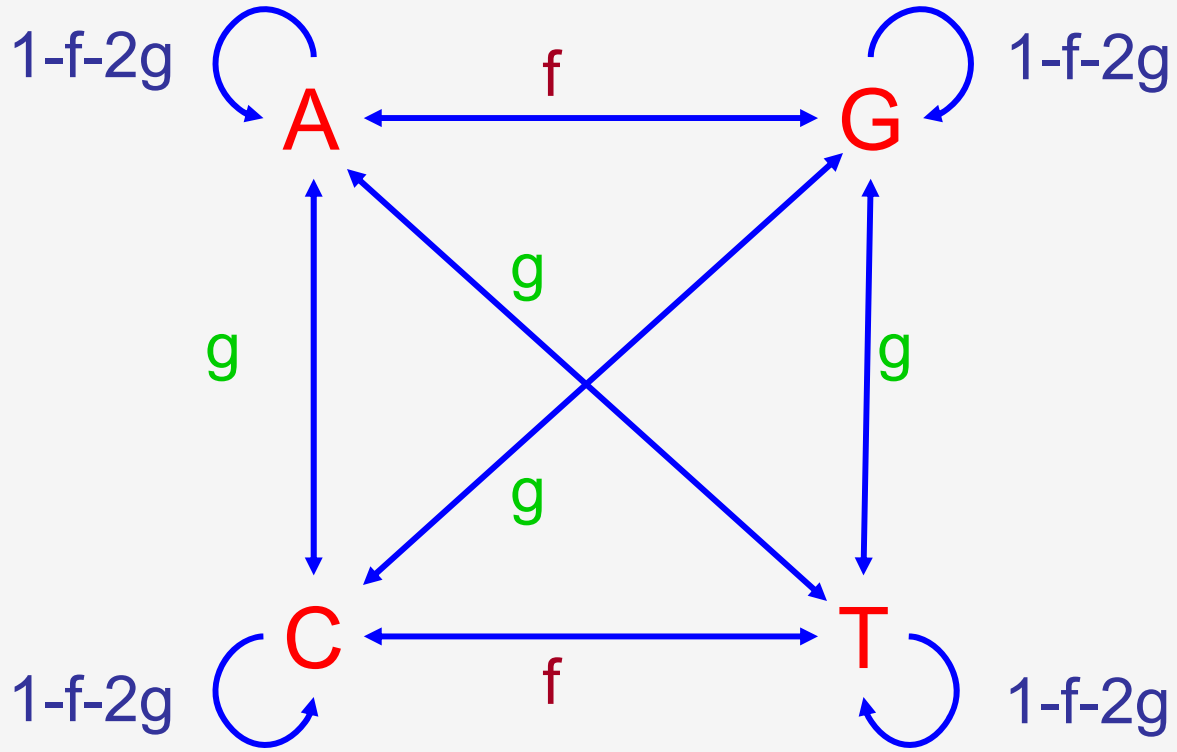
	hum	mac	bov	pla	chi	sha
hum	----	5	17	26	37	109
mac	5	----	19	26	36	100
bov	16	17	----	32	47	109
pla	23	23	27	----	35	106
chi	31	30	37	29	----	98
sha	65	62	65	64	61	----

$$K = \frac{-19}{20}\ln\left(1 - \frac{20}{19}d\right);$$

$$K * 100$$

Other models

- **Kimura's K2P model (1980)**



- **And others: F84 model (Felsenstein)..**

The general time-reversible model

- And there is of course the **general 12-parameter model** which has arbitrary rates for each of the 12 possible changes (from each of the 4 nucleotides to each of the 3 others).
- (Neither of these has formulas for the transition probabilities, but those can be done numerically.)

	A	C	G	T
A		$\alpha\pi_C$	$\beta\pi_G$	$\gamma\pi_T$
C	$\alpha\pi_A$		$\delta\pi_G$	$\varepsilon\pi_T$
G	$\beta\pi_A$	$\delta\pi_C$		$\nu\pi_T$
T	$\gamma\pi_A$	$\varepsilon\pi_C$	$\nu\pi_G$	

Concluding points on theoretical models of substitutions

- Most models assume that sites evolve independently (which is not entirely realistic).
- more realistic models ? the more complicated the model, it is hard to compute the probabilities
- For proteins each of the transition probabilities are widely different, simple modeling not possible! – What to do? – Numerical Solution

Blackboard notes

- Re-derive JC based on rate matrix
- Introduce more complex model such as Kimura
- Use Kimura model to calculate transition/transversion rate ratio R

Evolution: Neutral or Selection based?

- Selectionists: evolutionary changes are due to natural selection
- Neutralists : most evolutionary change at the molecular level is driven by random drift of selectively neutral mutants rather than natural selection (Neutral Model)

Selectionists vs neutralists – agree to not agree!
- still not resolved; leaning towards selectionists

Problems with Neutral Theory

- Neutral mutations, should vary as a function of generation/life time; there should be more changes per million years for rodents than for primates
- However, some proteins evolve at constant rate among lineages without regard to generation time: same rate between mice/rats and chimp/human

Molecular Clock Hypothesis

For any given protein or DNA sequence, the rate of evolution is approximately constant over time in all evolutionary lineages

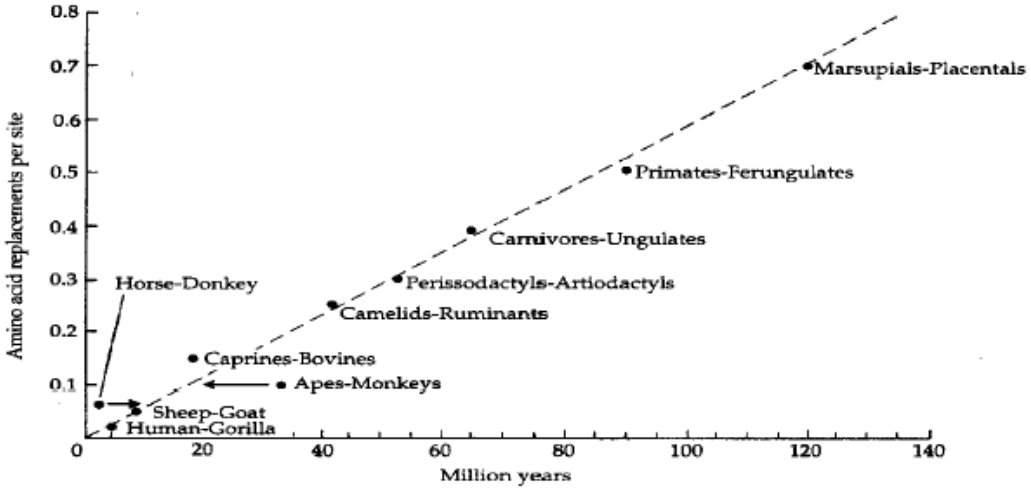
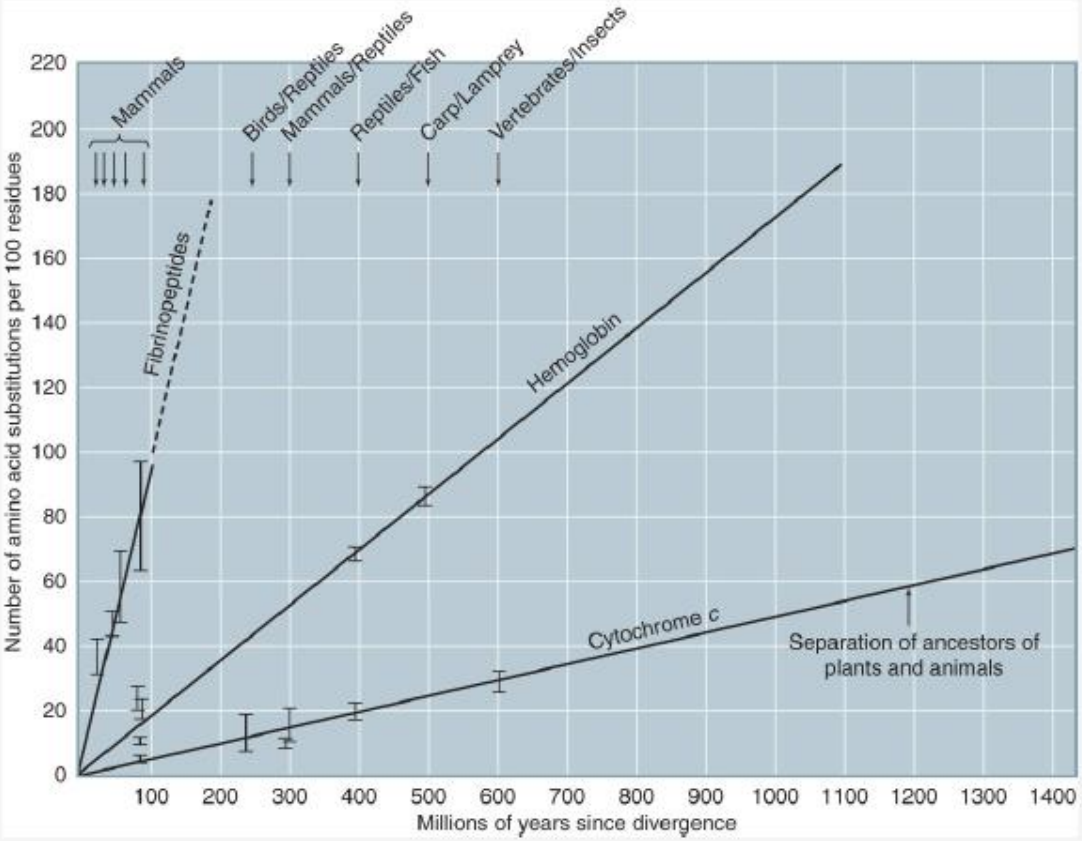


FIGURE 4.15 Number of amino acid replacements per amino acid site in a combined sequence consisting of hemoglobins α and β, cytochrome *c*, and fibrinopeptide A among various mammalian groups plotted against geological estimates of divergence times. The dashed line represents the molecular clock expectation 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).

Linus Pauling

Emile Zuckerkandl

Different proteins have different rates



Next week: Phylogenetic trees