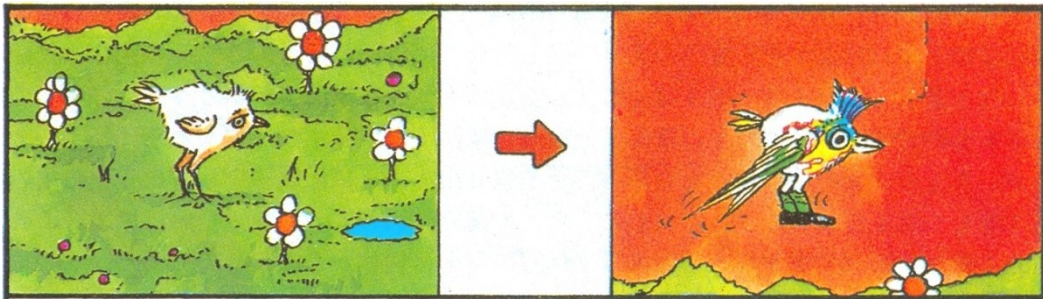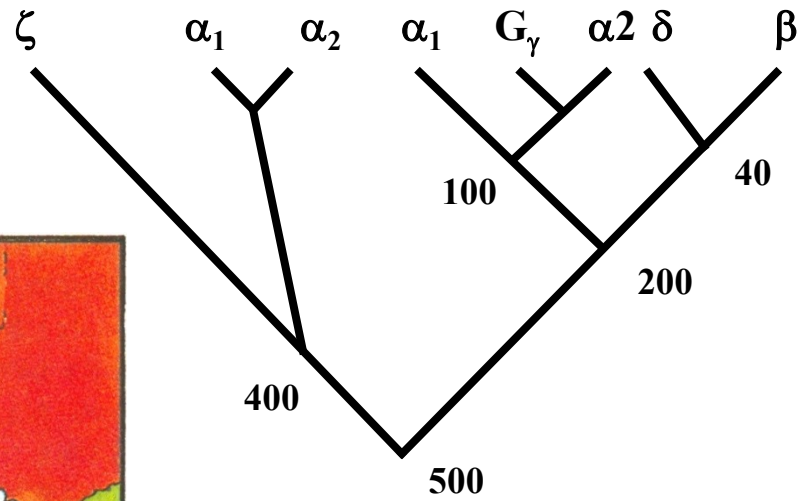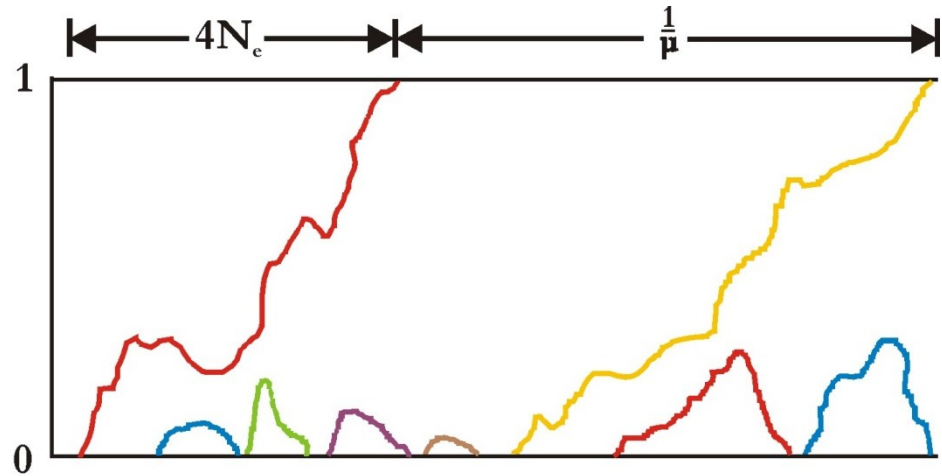# MOLECULAR EVOLUTION

# Substitutional load and the cost of natural selection

substitution = replacement of one allele by other (tj. fixation of a new allele)

if an individual do not reproduce, we call it <u>genetic death</u>

J.B.S. Haldane (1957):

advantageous mutation $\rightarrow$ fixation and replacement of a deleterious allele

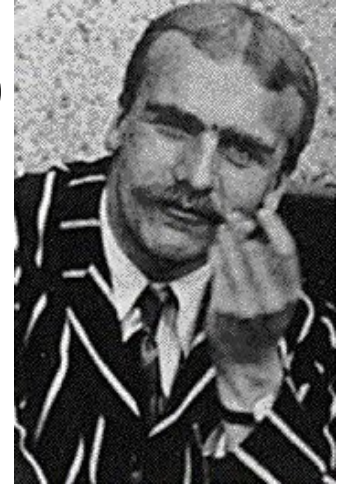as long as the original allele exists in the population, mean fitness is lower than maximal fitness

J.B.S. Haldane
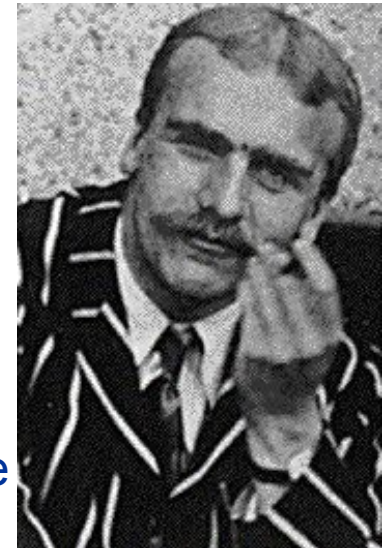
substitutional load[*)]: $L = 1 - \bar{w}$; when $\bar{w} = w_{max}$   $L = 0$

in general

$$L = \frac{w_{max} - \bar{w}}{w_{max}}$$

it measures to what extent an average member of the population is less fit than the best fit genotype

it expresses probability that an average individual dies before his/her reproduction

*) in general = genetic load; other loads: mutational: emergence of a deleterious allele; segregational: cost of homozygotes under overdominance

# Cost of natural selection:

We can envisage replacement of one allele in a population by another as a „selective death" of the original allele

The higher is the strength of selection, the more of the original (less advantageous) alleles are removed from the population (they „die")

If the natural selection was too strong it could cause extinction of the whole population $\Rightarrow$ overproduction of the offspring necessary!

eg. if the ratio of non-surviving to surviving alleles is 0,1/0,9, each survivor have to produce by 1/9 more descendants,
but if the ratio is 0,999/0,001 $\rightarrow$ by ~1000 more descendants!

Haldane: upper limit of the cost of natural selection $\approx$ substitution of 1 gene per 300 generations

$\Rightarrow$ evolution cannot run too fast, the cost of selection would be too high

64 codons

20 amino acids

**Second position**

| | U | C | A | G | |
|---|---|---|---|---|---|
| **U** | UUU UUC *phe*<br>UUA UUG | UCU UCC UCA UCG *ser* | UAU UAC *tyr*<br>UAA *Stop*<br>UAG *Stop* | UGU UGC *cys*<br>UGA *Stop*<br>UGG *trp* | U C A G |
| **C** | CUU CUC CUA CUG *leu* | CCU CCC CCA CCG *pro* | CAU CAC *his*<br>CAA CAG *gln* | CGU CGC CGA CGG *arg* | U C A G |
| **A** | AUU AUC *ile*<br>AUA<br>AUG *met* | ACU ACC ACA ACG *thr* | AAU AAC *asn*<br>AAA AAG *lys* | AGU AGC *ser*<br>AGA AGG *arg* | U C A G |
| **G** | GUU GUC GUA GUG *val* | GCU GCC GCA GCG *ala* | GAU GAC *asp*<br>GAA GAG *glu* | GGU GGC GGA GGG *gly* | U C A G |

First position (5′-end)

Third position (3′-end)

Initiation   Termination

excess of synonymous nucleotide substitutions → esp. at the 3rd position

M. Kimura (1977): mRNA sequences of humans and rabbits →
of 53 nucleotide positions 6 differences, of them only 1 nonsynonymous

× theoretically only 24% of differences should be synonymous

likewise M. Grunstein (1976):

evolutionary rate of Histone H4 in 2 sea urchin species
84 bp mtDNA → 9 of 10 differences synonymous
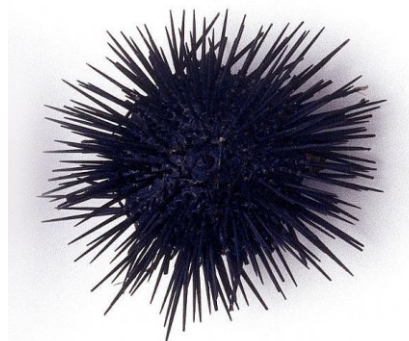
# NEUTRAL THEORY OF MOLECULAR EVOLUTION

Modern Synthesis: selection *vs*. drift debate

beginning of the 1960s $\rightarrow$ amino acid sequences in proteins

1966: protein elektrophoresis

Richard Lewontin and Jack Hubby - *Drosophila pseudoobscura*;

Harry Harris - humans

$\rightarrow$ high level of polymorphism

Rate of molecular evolution is too high

Genetic variation in natural populations is too high

... both would require too high cost of natural selection $\Rightarrow$ polymorphism cannot be maintained by selection

Rate of evolution at the molecular level is constant

Higher evolutionary rate in functionally less important parts of the molecule, in noncoding regions and pseudogenes

# Why so high polymorphism in populations?

Motoo Kimura: because alleles are selectively neutral, it lasts many
generations till a new mutation is fixed – meanwhile the population
must be polymorphic = <u>transient polymorphism</u>

During the process of fixation often a new allele appears by mutation $\Rightarrow$
in a sufficiently large population at each point in time there is a large
amount of variation
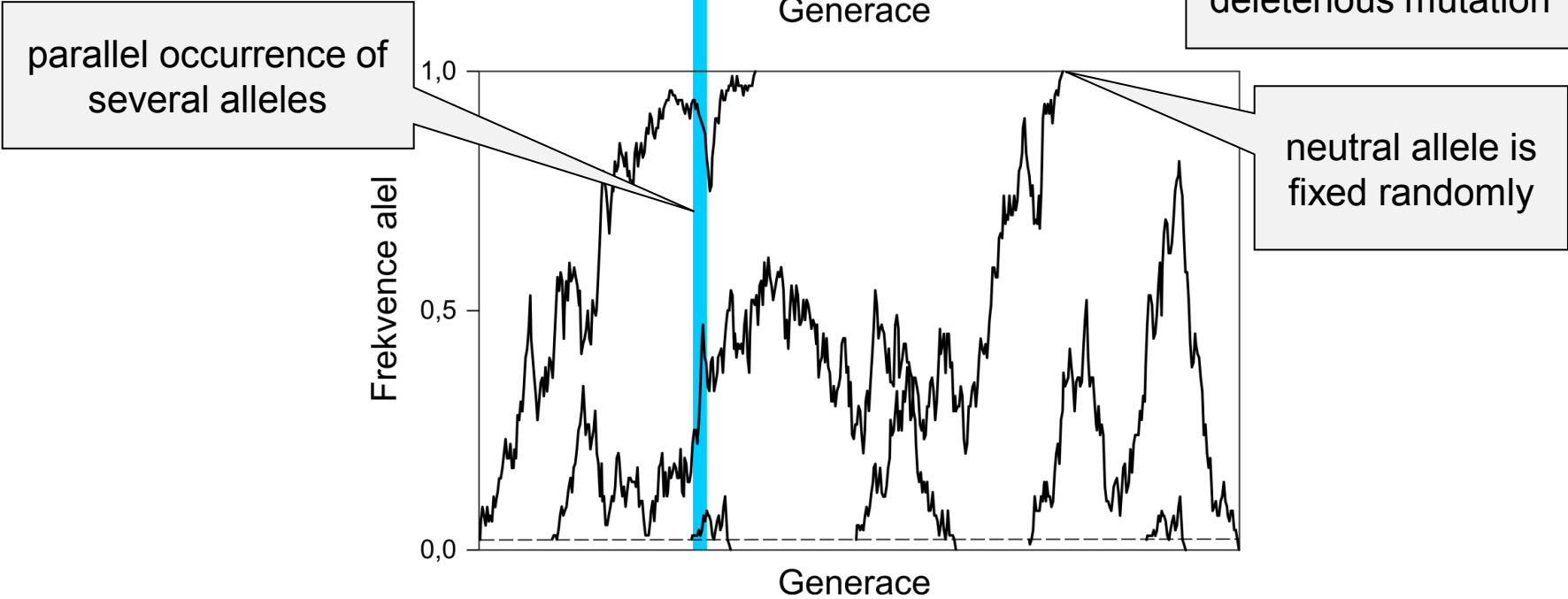
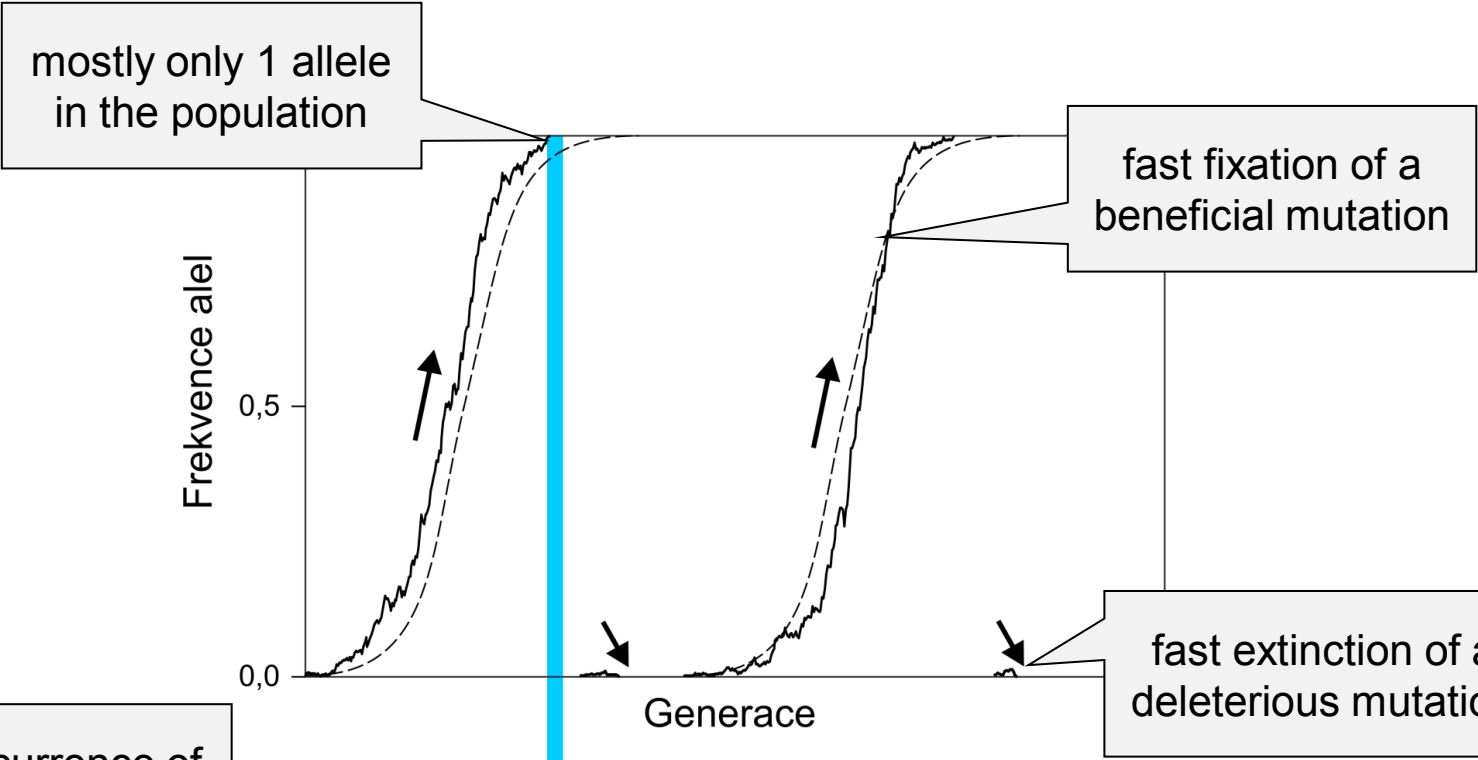Population is in equilibrium between drift
and mutation

M. Kimura

M. Kimura (1968)

J.L. King & T.H. Jukes (1969) $\longrightarrow$ neutral theory of
molecular evolution

mostly only 1 allele in the population

fast fixation of a beneficial mutation

fast extinction of a deleterious mutation

parallel occurrence of several alleles

neutral allele is fixed randomly

Frekvence alel
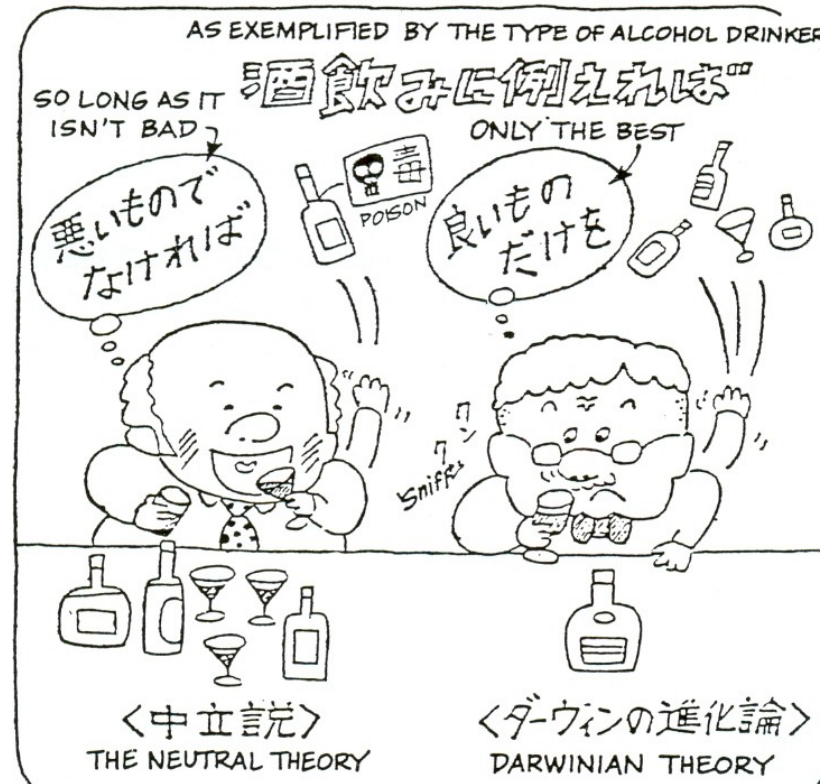
0,5

0,0

Generace

Frekvence alel

1,0

0,5

0,0

Generace

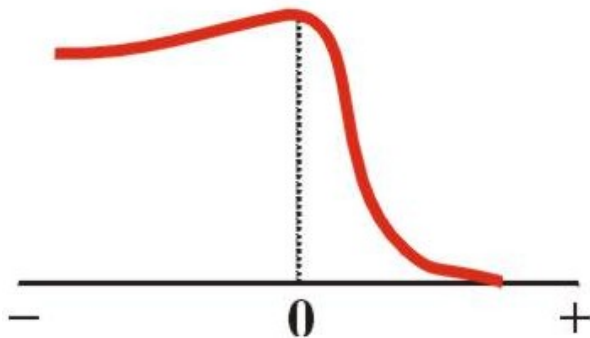# Basic principles of the neutral theory:

## 1. most allele substitutions in a population are neutral ($\Rightarrow$ drift)
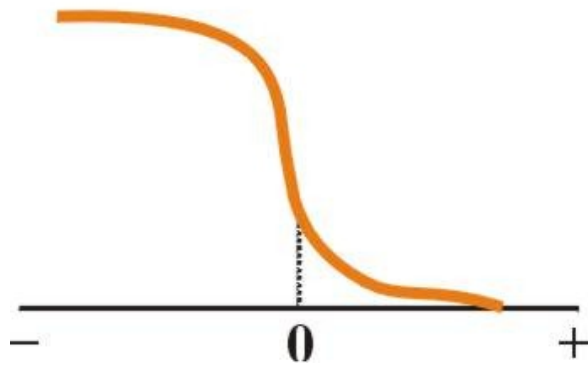
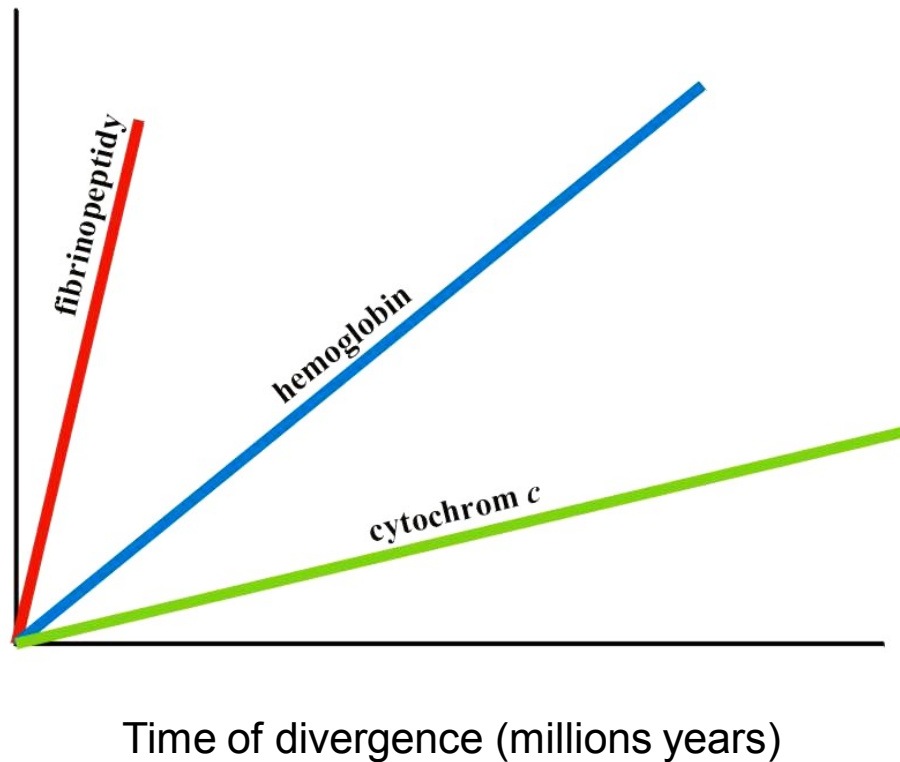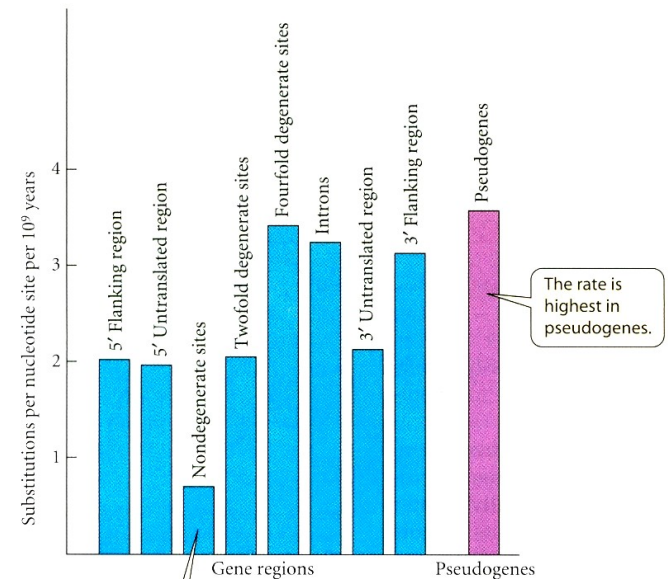# 2. evolutionary rates in differently important proteins are different

| | |
|---|---|
| fibrinopeptides | 8,3 |
| pancreatic ribonuclease | 2,1 |
| lysozyme | 2,0 |
| alpha-globin | 1,2 |
| insulin | 0,44 |
| cytochrome *c* | 0,3 |
| histone H4 | 0,01 |

No. AA substitutions per 100 molecules

fibrinopeptidy

hemoglobin

cytochrom c

Time of divergence (millions years)

Substitutions per nucleotide site per $10^9$ years

5' Flanking region
5' Untranslated region
Nondegenerate sites
Twofold degenerate sites
Fourfold degenerate sites
Introns
3' Untranslated region
3' Flanking region
Pseudogenes

The rate is highest in pseudogenes.

Gene regions

Pseudogenes

Nucleotide substitution rates are lowest at nondegenerate nucleotide positions, in which any mutation alters the amino acid specified.

# 3. diverse evolutionary rate in different parts of proteins (binding sites × structural areas)

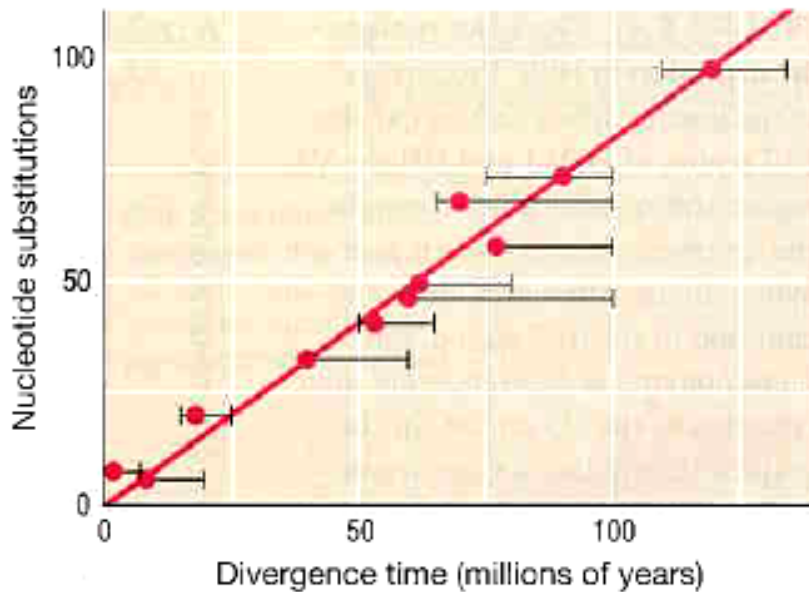Eg.: transient receptor potential vanilloid (TRPV) channel protein:



ATP molecule

N

C

Conserved      Divergent

binding sites are more conserved
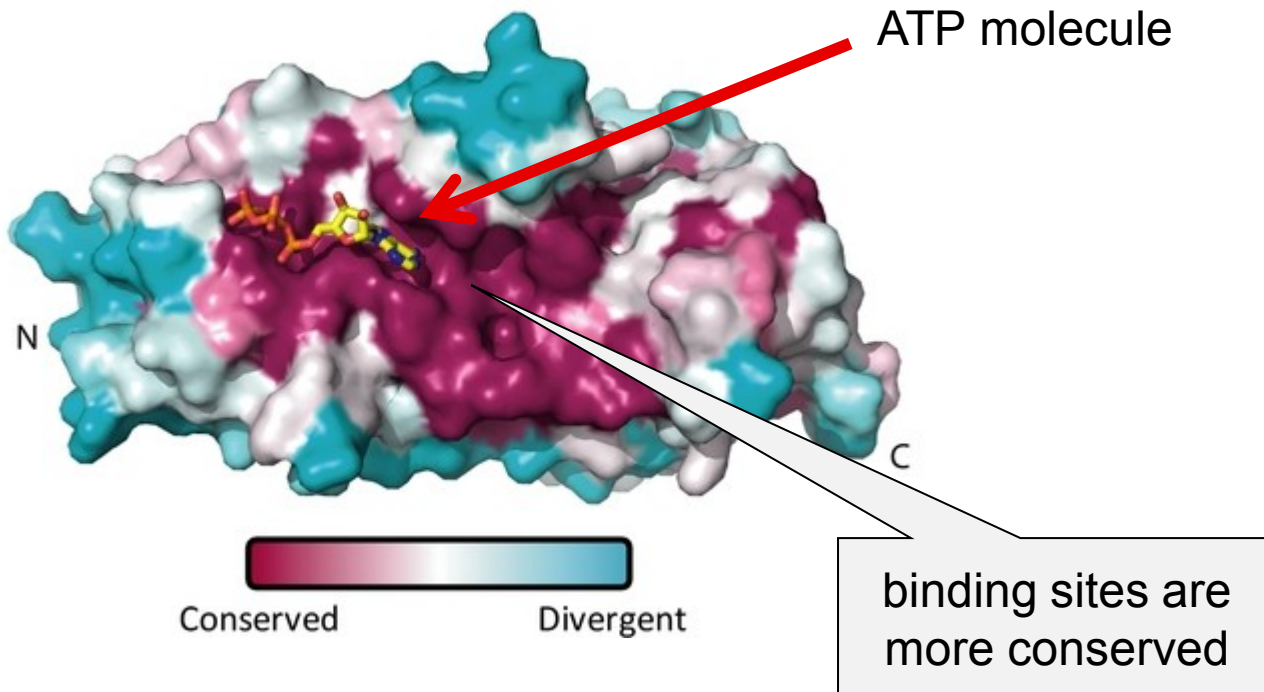
# 4. different evolutionary rates at individual codone positions

**Table 4. Relative frequencies of different types of mutational substitutions in a random protein-coding sequence.**

| Substitution | Number | Percent |
|---|---|---|
| Total in all codons | 549 | 100 |
| Synonymous | 134 | 25 |
| Nonsynonymous | 415 | 75 |
| Missense | 392 | 71 |
| Nonsense | 23 | 4 |
| Total in first position | 183 | 100 |
| | | |
| Missense | 166 | 91 |
| Nonsense | 9 | 5 |
| Total in second position | 183 | 100 |
| | | |
| Missense | 176 | 96 |
| Nonsense | 7 | 4 |
| Total in third position | 183 | 100 |
| | | |
| Missense | 50 | 27 |
| Nonsense | 7 | 4 |

# 5. evolutionary rate of a given protein in various species roughly constant

Kimura (1983), vertebrates, $\alpha$-globin:

Wilson (1977), mammals, 7 proteins:
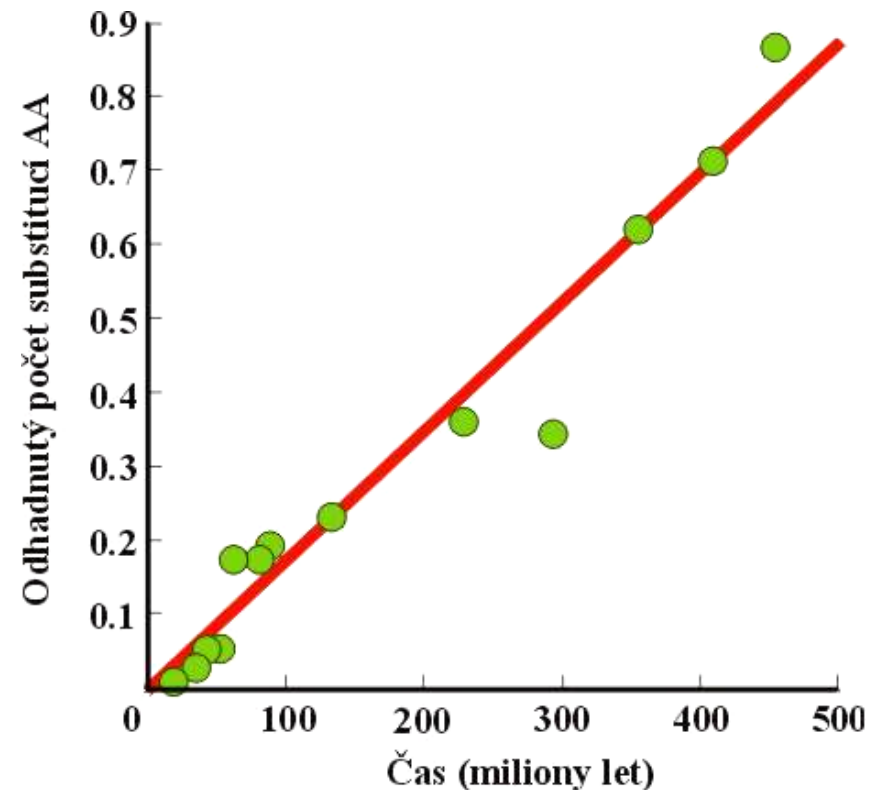
mostly does not concern morphological, physiological, and behavioural trais

cannot explain adaptations

many deleterious mutations, however, these rapidly removed by selection

selection acts also at the molecular level but most mutations have only small effects on fitness $\Rightarrow$ important role of drift

# Haldane s cost of selection was overestimated:

selection mostly <u>soft</u>, not hard

<u>frequency-dependent selection</u> rather than overdominance

selection does not affect individual loci independently (epistasis)
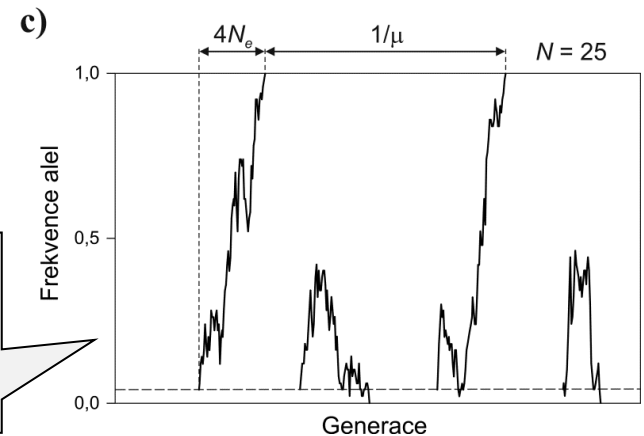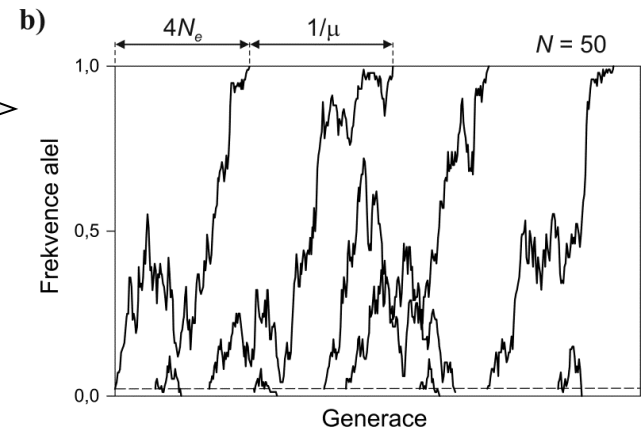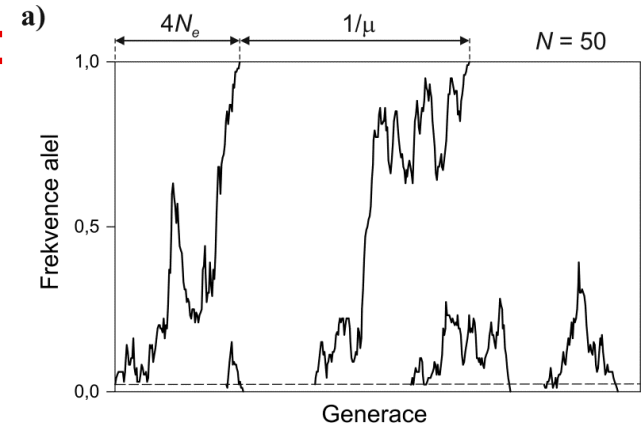
# Theoretical principles of neutral theory:

Mean time to fixation of a novel mutation
$= 4N_e$

Mean interval between fixations
$= 1/\mu$

In small populations faster fixations but
longer interval between them:

moderately sized population: mutations more frequent

small population: mutations infrequent

a) $4N_e$  $1/\mu$  $N = 50$
Frekvence alel
1,0
0,5
0,0
Generace

b) $4N_e$  $1/\mu$  $N = 50$
Frekvence alel
1,0
0,5
0,0
Generace

c) $4N_e$  $1/\mu$  $N = 25$
Frekvence alel
1,0
0,5
0,0
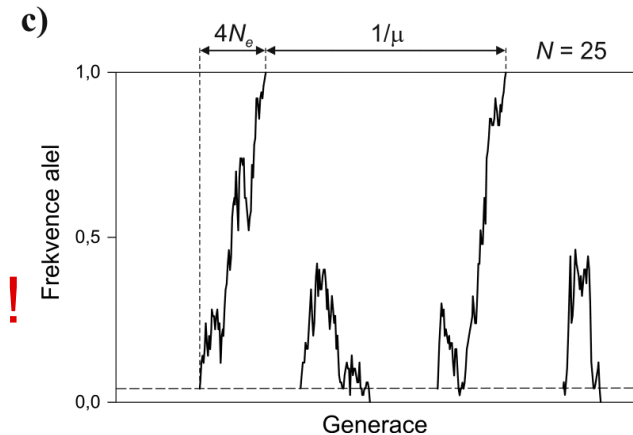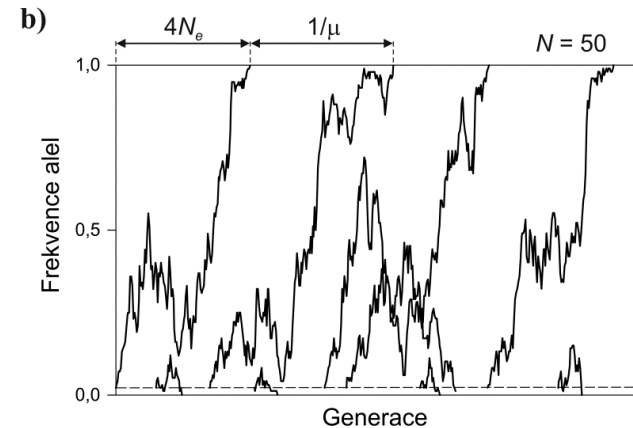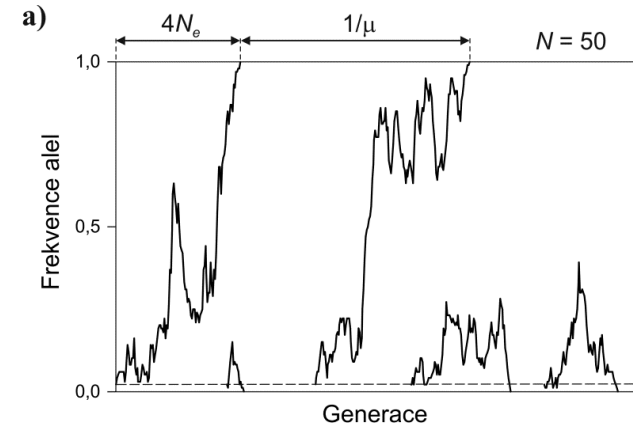Generace

# Theoretical principles of neutral theory:

Mean time to fixation of a novel mutation
$$= 4N_e$$

Mean interval between fixations
$$= 1/\mu$$

Frequency of substitutions (replacements of one allele by another in populations):

$$1/(2N_e) \times 2N_e\mu = \underline{\mu}$$

$\Rightarrow$ rate of neutral evolution is independent of $N_e$, depends only on frequency of neutral mutations $\mu$ !



a) $N = 50$, $4N_e$, $1/\mu$, axes: Frekvence alel, Generace

b) $N = 50$, $4N_e$, $1/\mu$, axes: Frekvence alel, Generace

c) $N = 25$, $4N_e$, $1/\mu$, axes: Frekvence alel, Generace

## Theoretical principles of neutral theory:

mean balanced heterozygosity:

$$\frac{\theta}{\theta+1}, \text{ where } \theta = 4N_e\mu$$

larger populations $\Rightarrow$ higher heterozygosity

continual emergence of new mutations $\Rightarrow$ increase of variation

$\times$ its erosion by drift

$\Rightarrow$ continual replacement of one allele by another

Equilibrium between mutation and drift $\Rightarrow$ polymorphism (contrary to the mutation-selection eq.) is transient
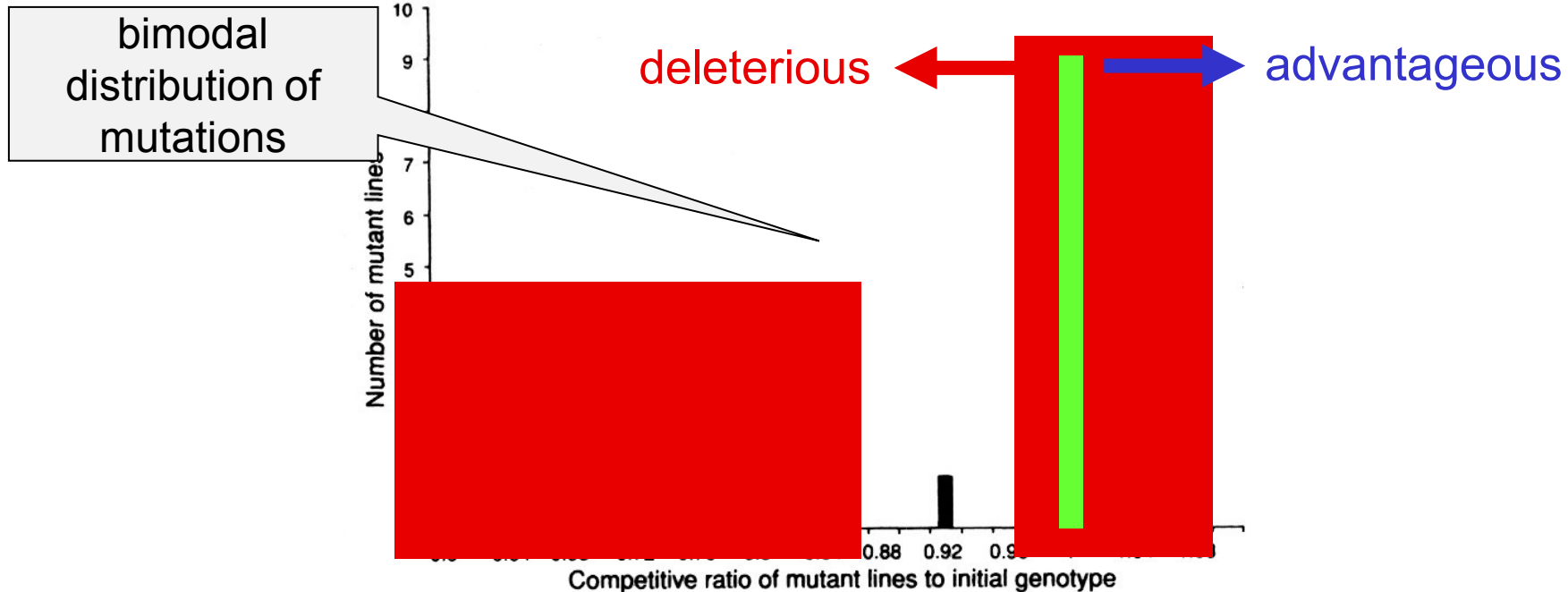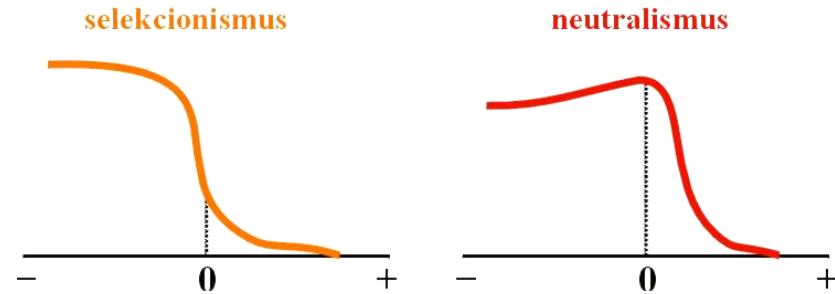
# Rate of neutral mutations:

Zeyl & DeVisser (2001):

yeast *Saccharomyces cerevisiae*

50 replicated populations,
    1 individual in each generation

the experiment does not detect extremely deleterious (lethal) mutations



selekcionismus    neutralismus

bimodal distribution of mutations

deleterious    advantageous

Number of mutant lines

0.88  0.92  0.9
Competitive ratio of mutant lines to initial genotype

# Test of neutral theory: range of heterozygosity
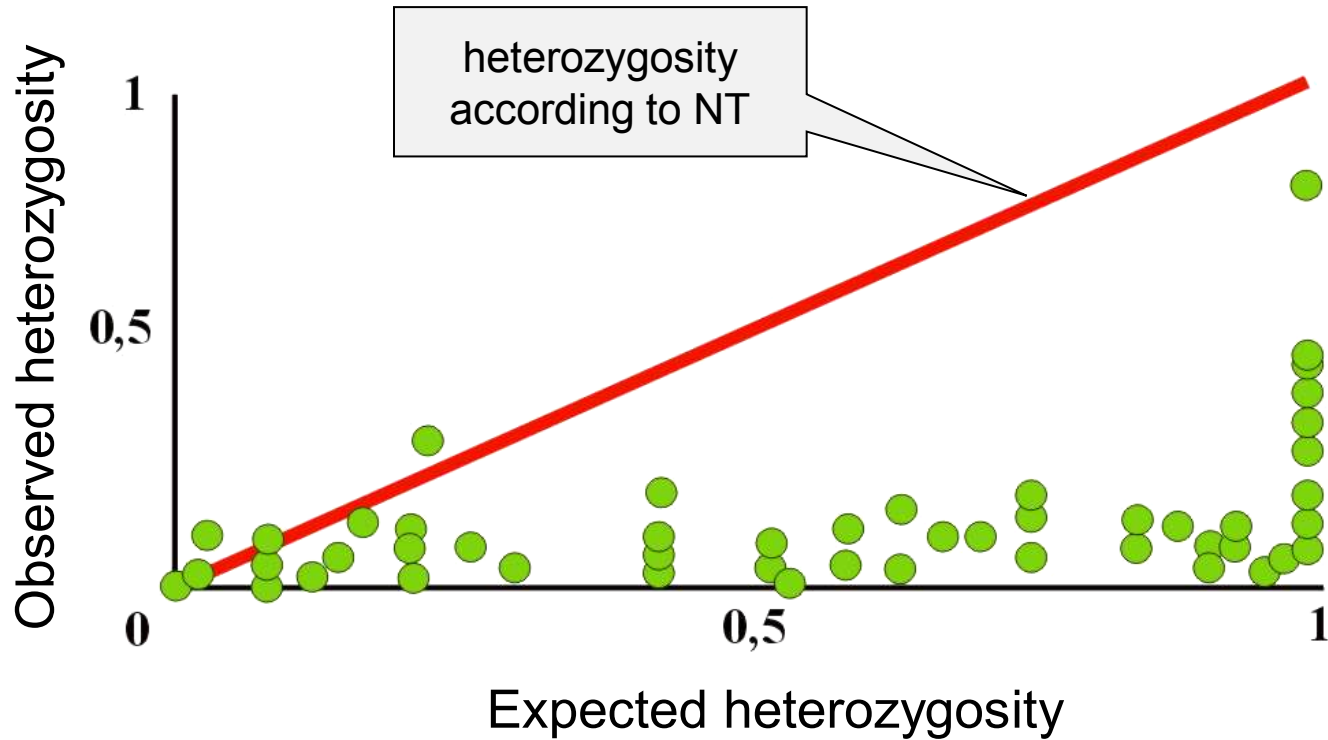


Observed <u>heterozygosity lower</u> than predicted by NT

# Test of neutral theory: range of heterozygosity
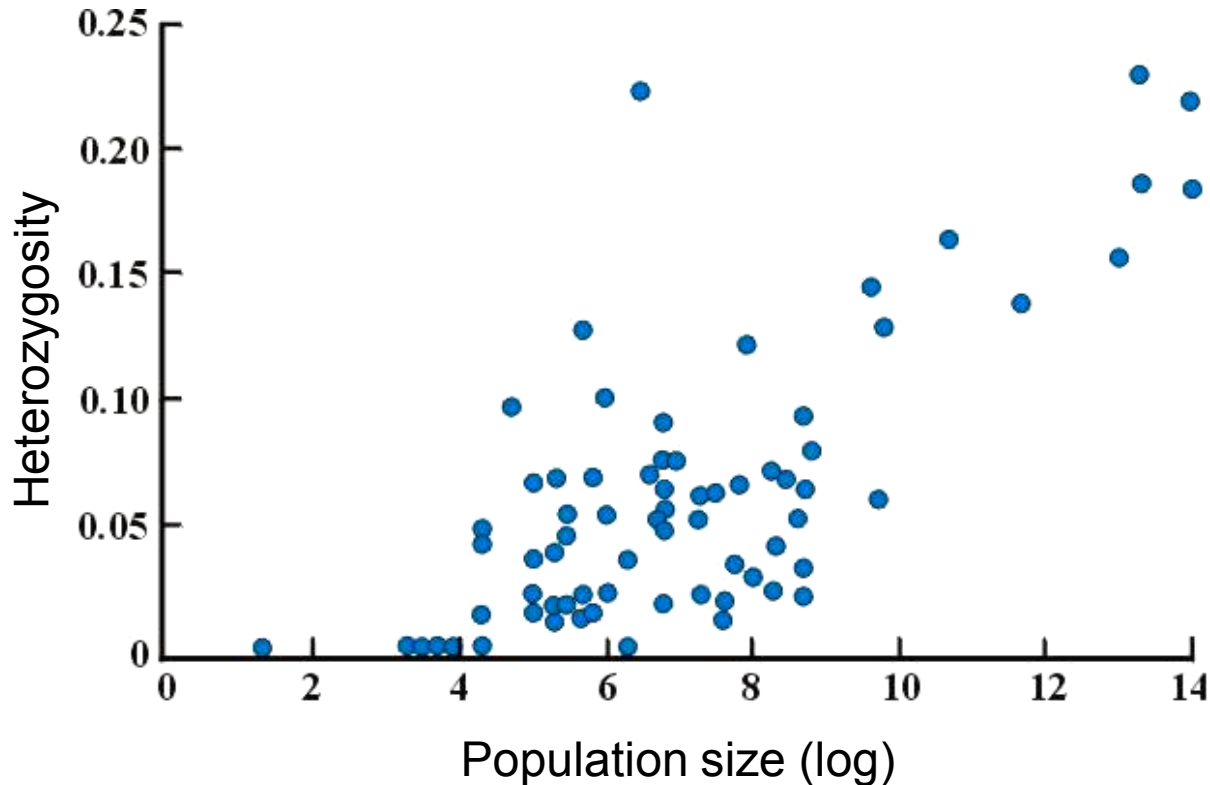


Given the enormous range of population sizes, the range of heterozygosities is too small

Tomoko Ohta tried to explain the deviations of the observed range of heterozygosity from predictions:
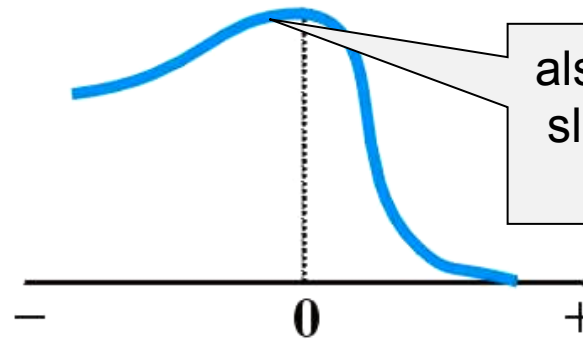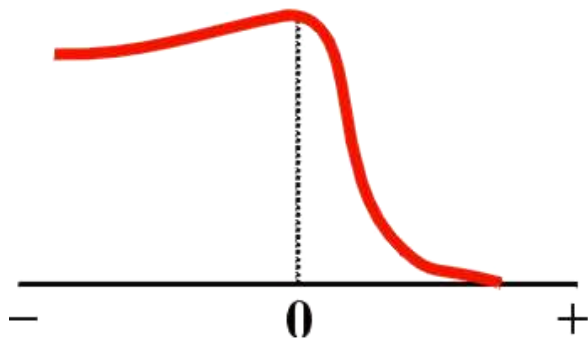
slightly deleterious mutations (SDM)

in small populations alleles behave as effectively neutral

**neutralism**

**slightly deleterious mutations**

also substitutions of slightly deleterious alleles

−     0     +

−     0     +

# Fixation probability of neutral, beneficial, and deleterious mutation:

Eg.:
What is the fixation probability of a mutation in a population of $N_e$ = 1000?

neutral mutation ($s$ = 0):                          $P$ = 0,05%

advantageous mutation ($s$ = 0,01):          $P$ = 20%

advantageous mutation ($s$ = 0,001):        $P$ = 2%

deleterious mutation ($s$ = -0,001):           $P$ = 0,004%

as $s \rightarrow 0$
higher
„neutrality"

We can conclude that

1) Not all advantageous mutations may be fixed in the population

2) Conversely, with low probability also deleterious mutations can be fixed

What is the fixation probability of a mutation in a population of $N_e$ = 10 000?

neutral mutation ($s$ = 0):                         $P$ = 0,005%
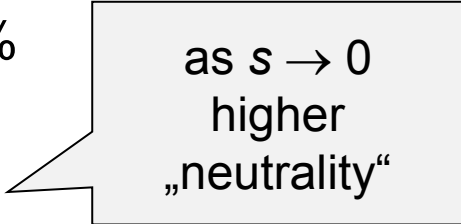
advantageous mutation ($s$ = 0,01):        $P$ = 20%

advantageous mutation ($s$ = 0,001):       $P$ = 2%

deleterious mutation ($s$ = -0,001):          $P$ = $2.10^{-17}$%

in a large population $P$ of an advantegous allele is the same as in a small one but for a deleterious allele $P \rightarrow 0$
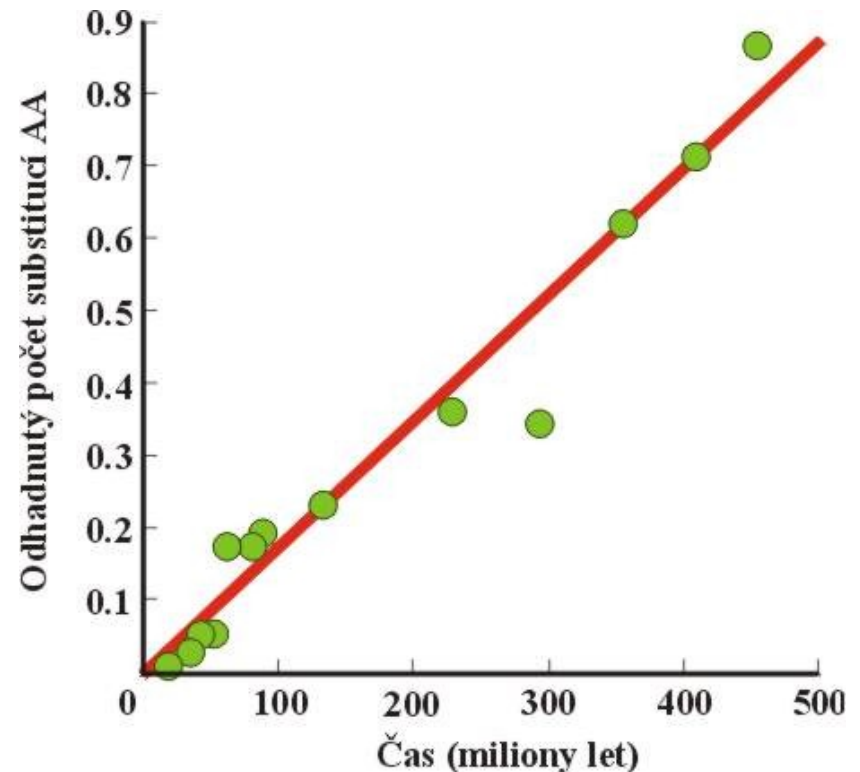
Conclusions:

1) In large populations selection plays a much more important role; conversely, with decreasing population size the role of drift is increasing

2) Harmfulness of a mutation is inversely proportional to population size: the more it approaches zero, the larger the population may be for allele fixation (drift exceeds negative selection) and vice versa: the stronger selection against an allele, the smaller the population must be to allow drift to play a substantial role

3) This means that in small populations slightly deleterious mutations behave as effectively neutral
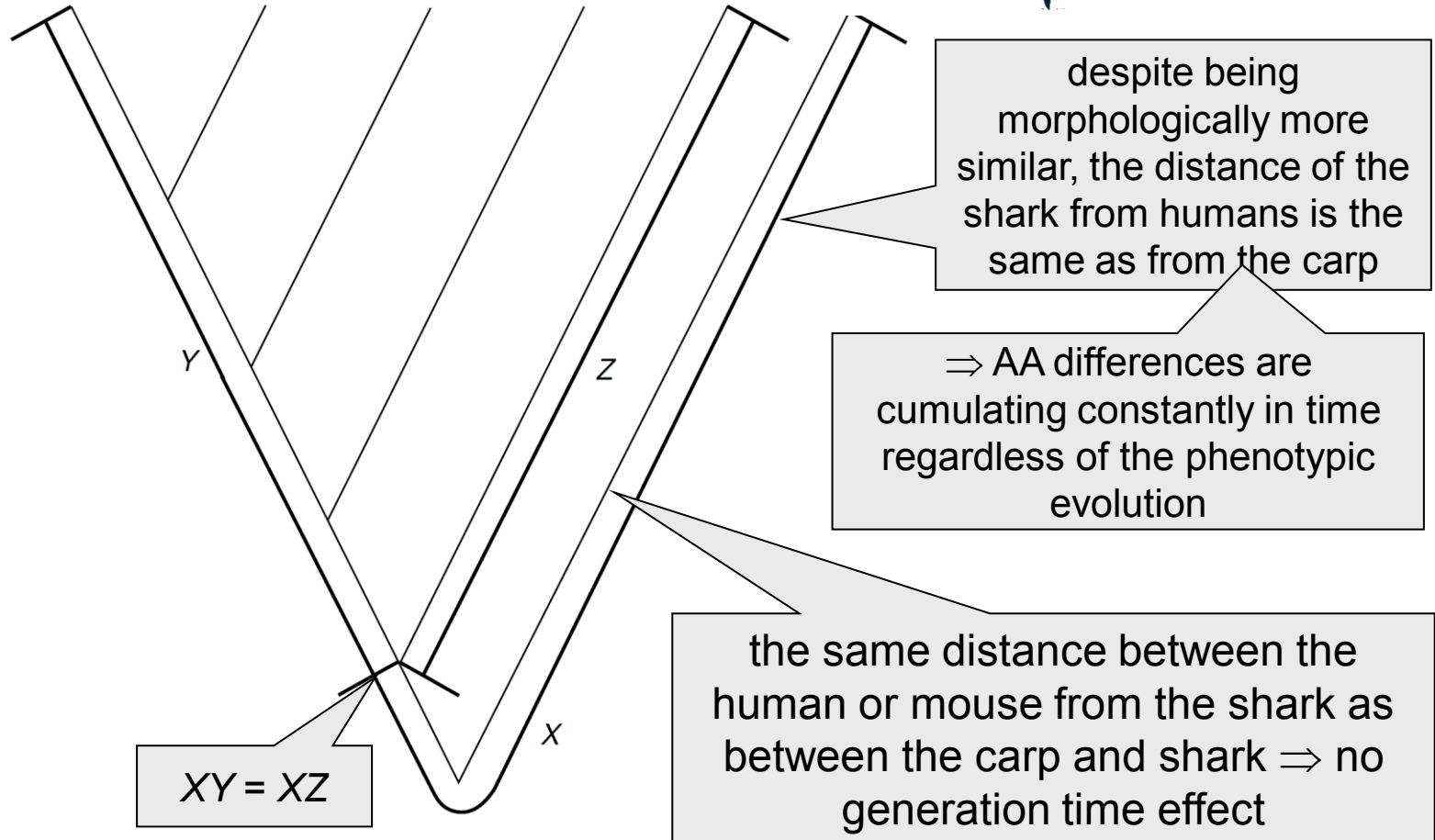
# MOLECULAR CLOCK

Zuckerkandl & Pauling (1962-65)

AA and/or nucleotide substitution rate is constant

effect of generation time:
dependency on absolute or
generation time?

# AA sequences of the α-chain of hemoglobin of 6 vertebrate species:



Human    Mouse    Chicken    Newt    Carp

*Y*

*Z*

*X*

$$XY = XZ$$

despite being morphologically more similar, the distance of the shark from humans is the same as from the carp

⇒ AA differences are cumulating constantly in time regardless of the phenotypic evolution

the same distance between the human or mouse from the shark as between the carp and shark ⇒ no generation time effect
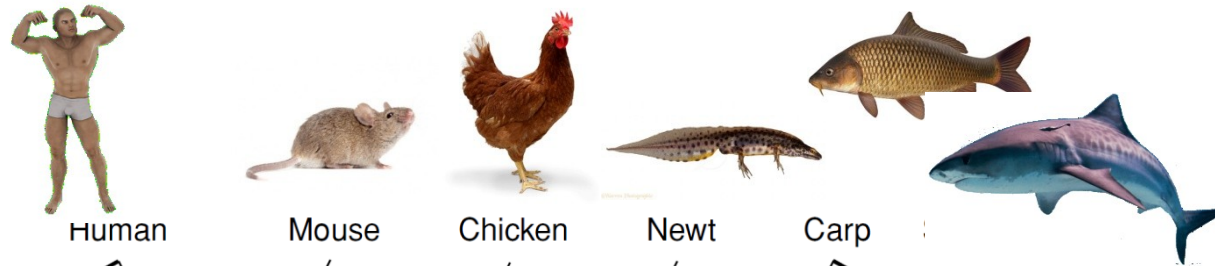
# Generation or absolute time?

Accumulation of neutral substitutions in placental mammals:
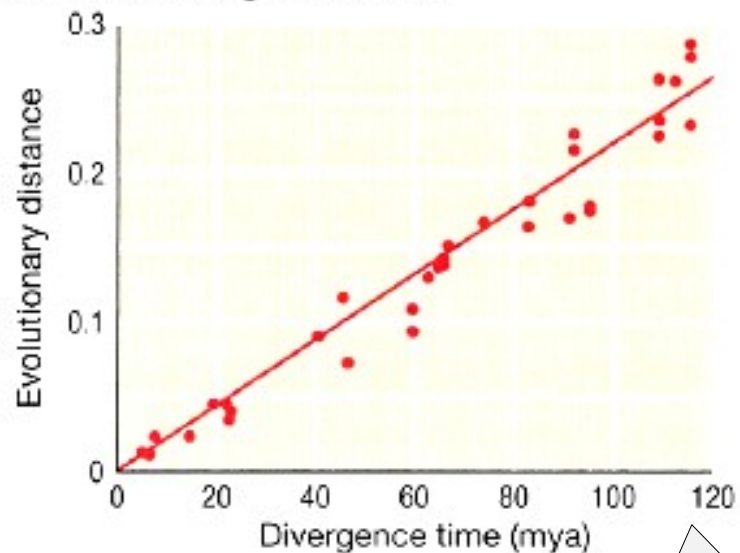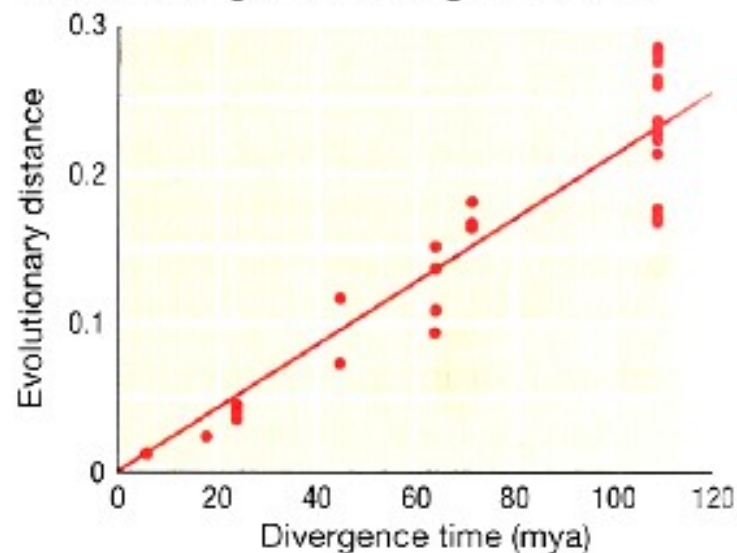


both dating methods show almost constant rate independent of generation time

# Population size and generation time:



species with larger populations tend to have shorter generation times
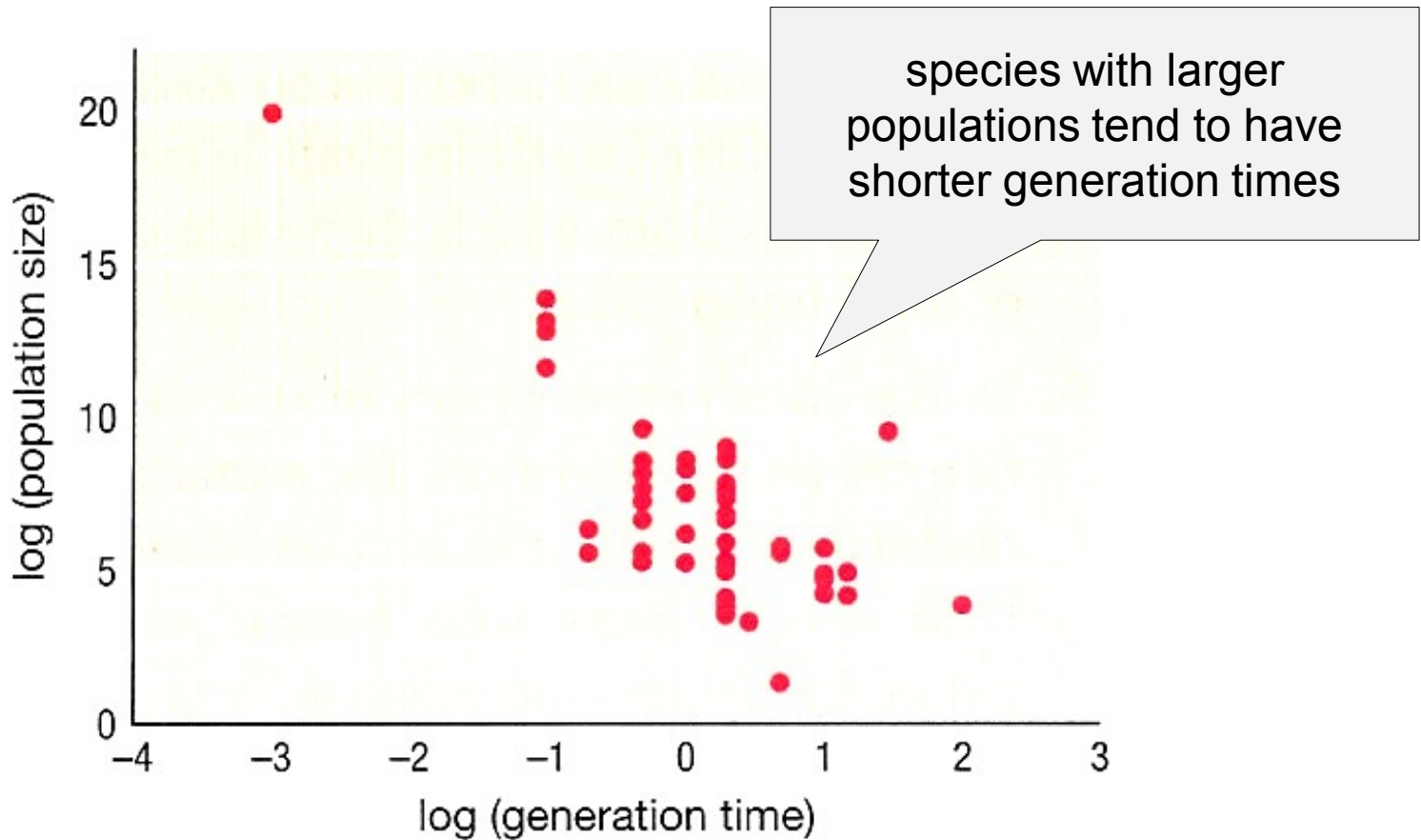
⇒ potential explanation of absolute time dependence: in smaller populations also slightly deleterious alleles are fixed

# But molecular clock does not „tick" with the same pace in different groups

eg. cetaceans < „artiodactyls"< primates < murine rodents
in primates Old World monkeys > „apes" > humans

# Problem of sequence saturation:
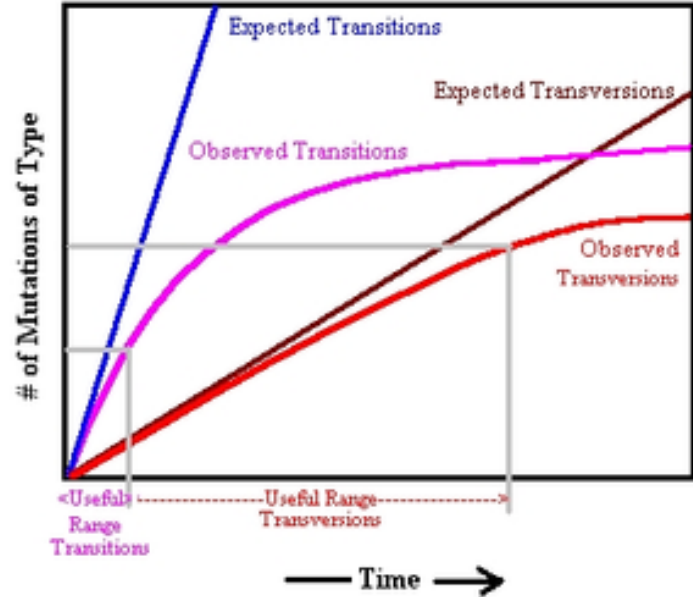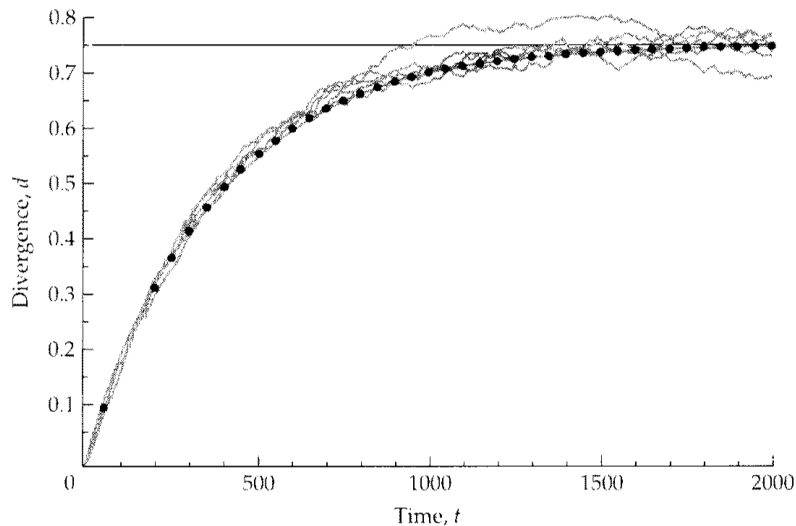


$\rightarrow$ using an appropriate evolutionary model („straightening" of the curve)
relaxed molecular clock method

# CONCERTED EVOLUTION AND MOLECULAR DRIVE

ribosomal DNA

globin genes



the ape species pairs differ by ~ 2,5 AA substitutions in the $\alpha 1$ i $\alpha 2$ gene ...

... only few differences accumulated between the $\alpha 1$ a $\alpha 2$ gene ...

... in fact this duplication is older than 300 million years

current genes have emerged through a series of duplications
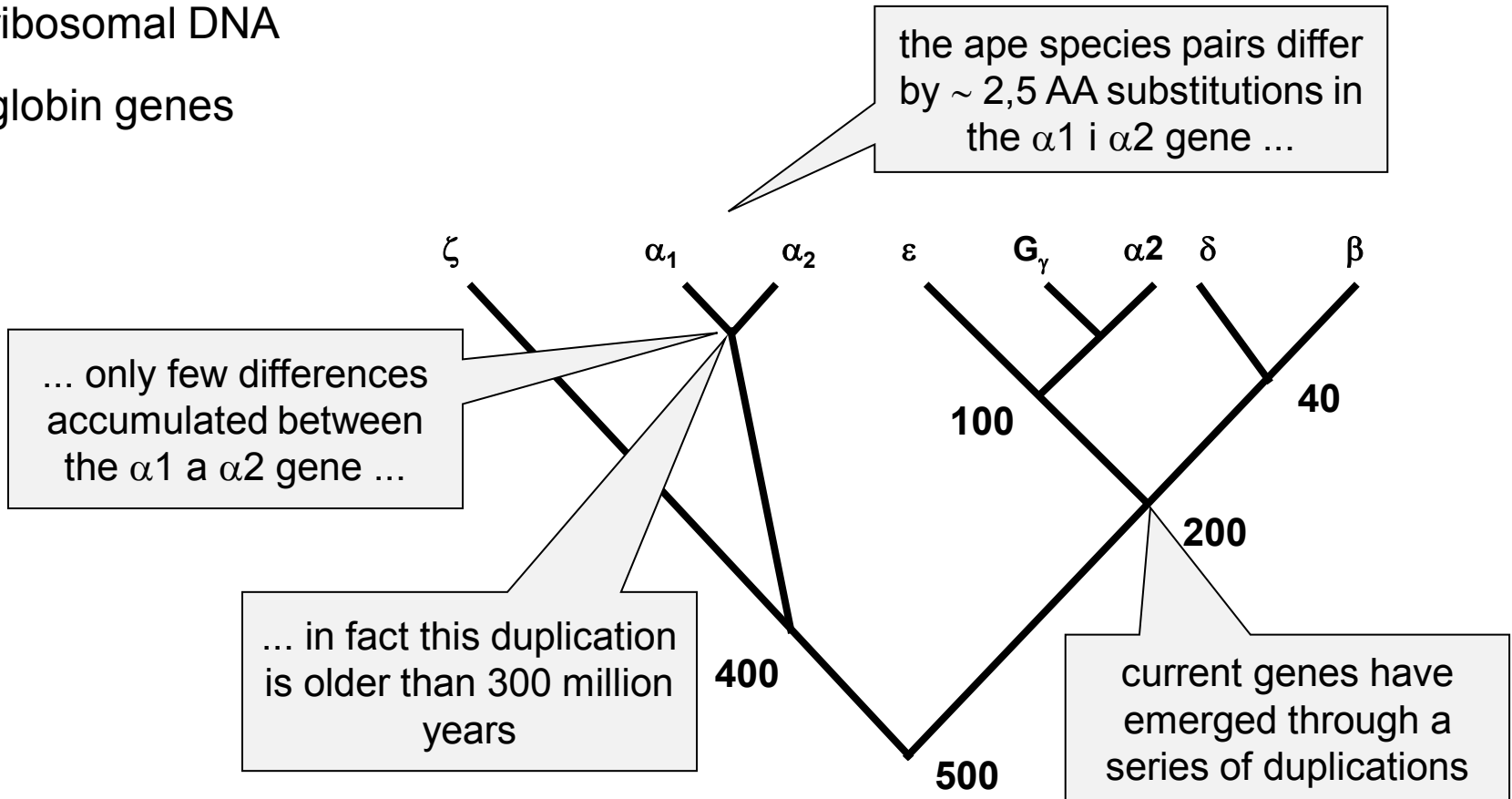
$\Rightarrow$molecular clock is invalid in this case, the genes do not evolve independently – the evolution is <u>concerted</u>

Gabriel Dover (1982): Molecular drive
mechanism different from selection and drift

Mechanisms of concerted evolution:

1. unequal crossing-over

gain of a mutant copy

loss of a normal copy

gain of a mutant copy

copy loss

copy gain

... etc. ...

# 2. replication slippage

## Figure Q-5: The Polymerase Slippage Model

### A) Slippage Event

template strand

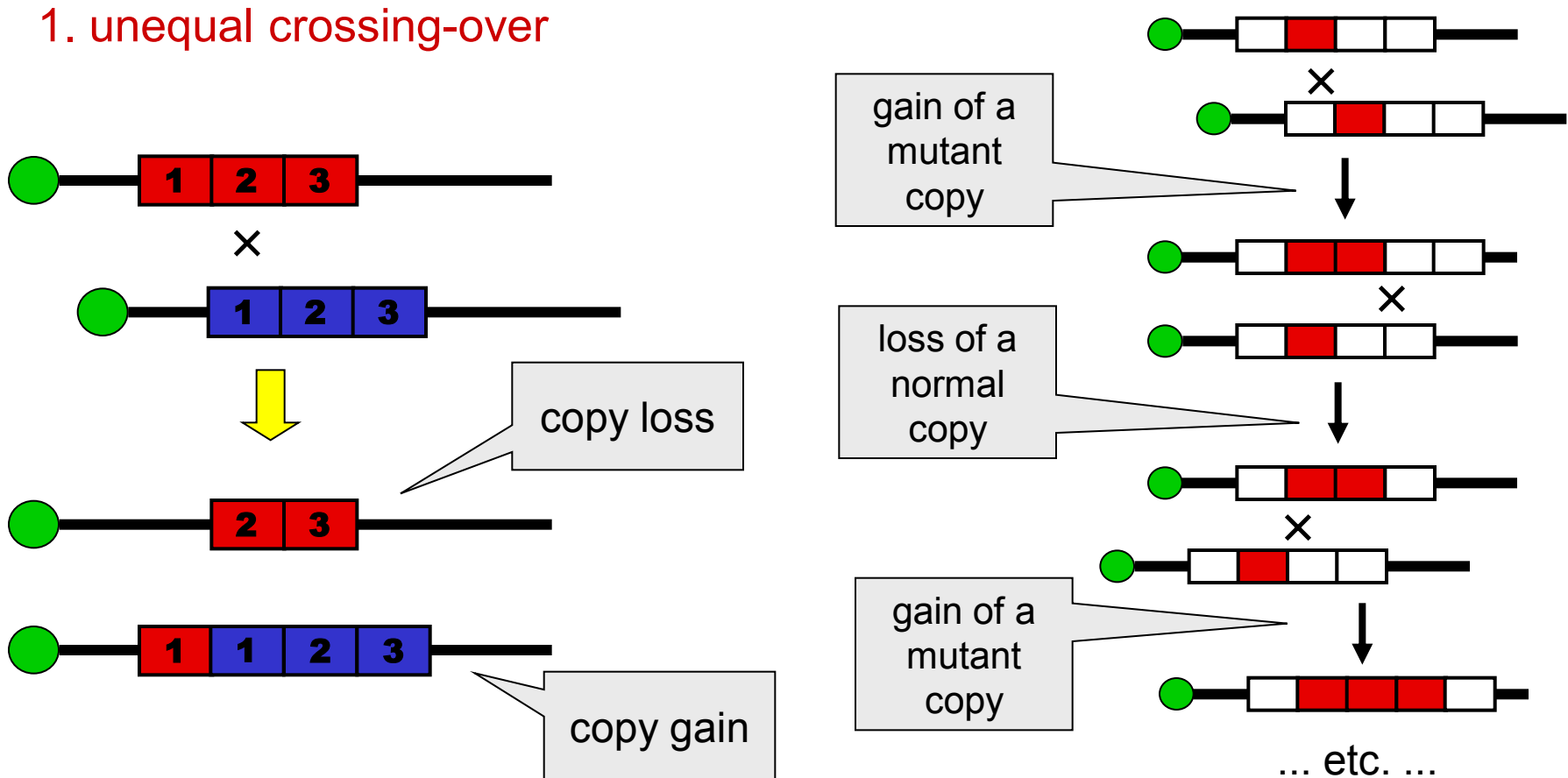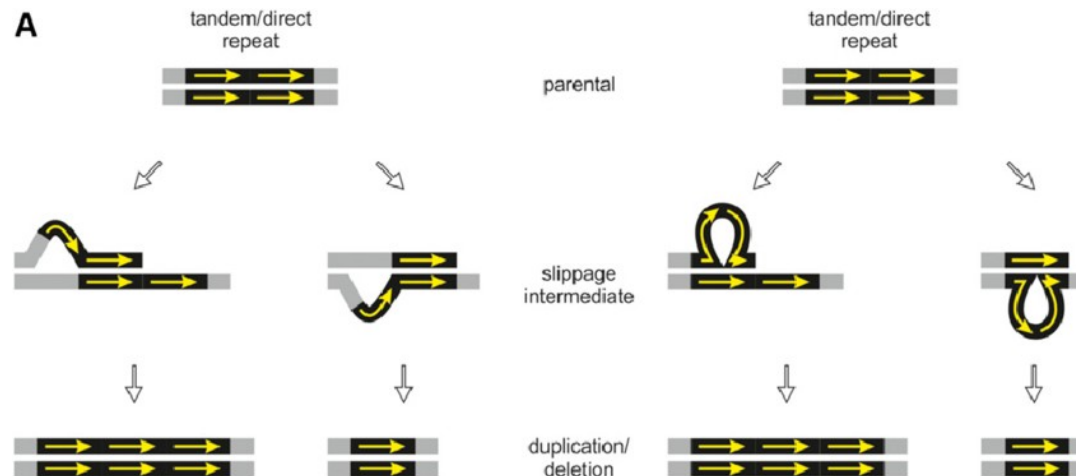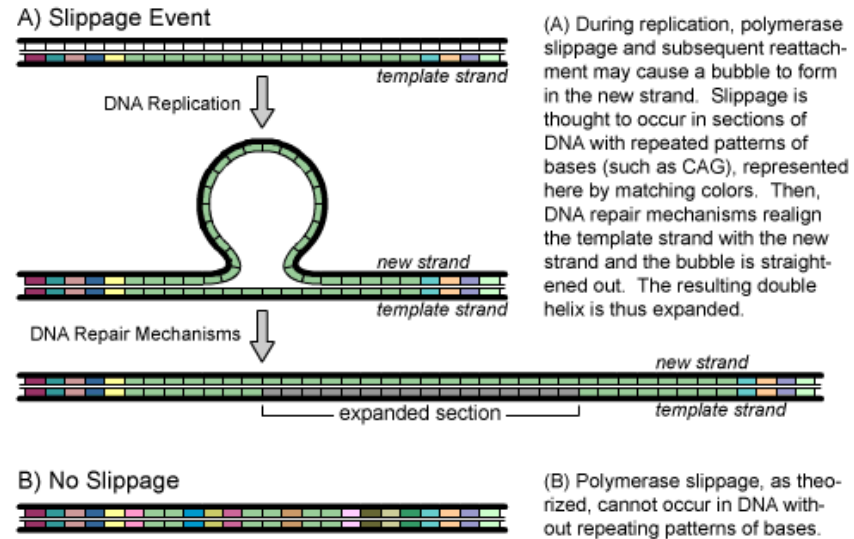DNA Replication

new strand

template strand

DNA Repair Mechanisms

new strand

expanded section

template strand

(A) During replication, polymerase slippage and subsequent reattachment may cause a bubble to form in the new strand. Slippage is thought to occur in sections of DNA with repeated patterns of bases (such as CAG), represented here by matching colors. Then, DNA repair mechanisms realign the template strand with the new strand and the bubble is straightened out. The resulting double helix is thus expanded.

### B) No Slippage

(B) Polymerase slippage, as theorized, cannot occur in DNA without repeating patterns of bases.

**A**

tandem/direct repeat

tandem/direct repeat

parental

slippage intermediate

duplication/ deletion

# 3. gene conversion



(b) Double strand break

Double holiday junction ↓

► Cleavage orientation 1
▽ Cleavage orientation 2

HJ resolution
Crossover          Non-crossover

Heteroduplex          Heteroduplex
Mismatch correction

4:4   4:4          4:4   2:6          2:6   4:4          2:6   2:6
Restoration Restoration Restoration Conversion Conversion Restoration Conversion Conversion

3:5
Average
restoration/conversion

Gene conversion          Crossover

# Conclusions:

a consequence of unequal crossing-over and slippage = <u>change of a copy number</u>

a consequence of unequal c-o and gene conversion = <u>sequence homogenization</u>