Natural selection on the molecular level

Fundamentals of molecular evolution

How DNA and protein sequences evolve?

Genetic variability in evolution

- Mutations
  - forming novel alleles
- Inversions
  - change gene order
  - may impede recombination and fix a haplotype
- Duplications
  - may involve genes or gene fragments
  - or entire chromosomes and genomes
  - the main source of evolutionary innovations
- Horizontal gene transfer
  - including symbiotic events

Substitutions

- Why transitions are more frequent?
  - there are more possible transversions
  - observed ts/tv ratio from ~2 (nDNA) to ~15 (human mtDNA)
  - exception – plant mtDNA
  - the ts/tv ratio varies significantly between lineages

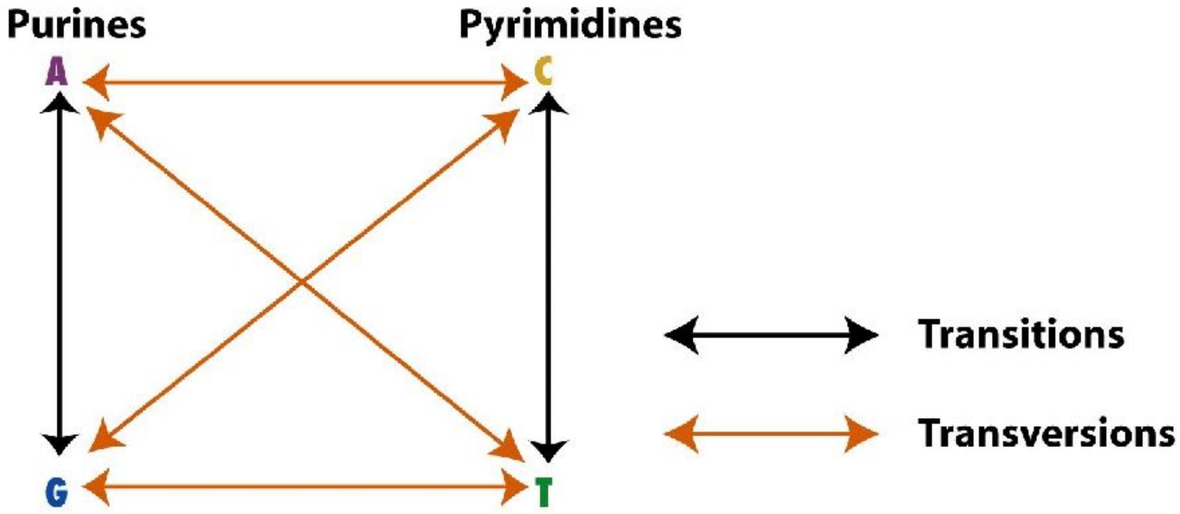


Figure 5-4 Evolutionary Analysis, 4/e

Transitions and transversions

- Why are the transitions more frequent?
  - Selection hypothesis (transitions are more likely to be silent and neutral)
  - But:
    - transitions are more frequent in rRNA genes, pseudogees and noncoding regions
    - transitions are more frequent in 4-fold degenerate positions (codons like CUN – Leu)

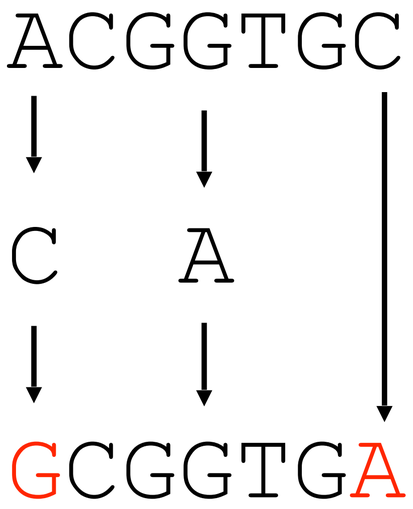
Transitions and transversions

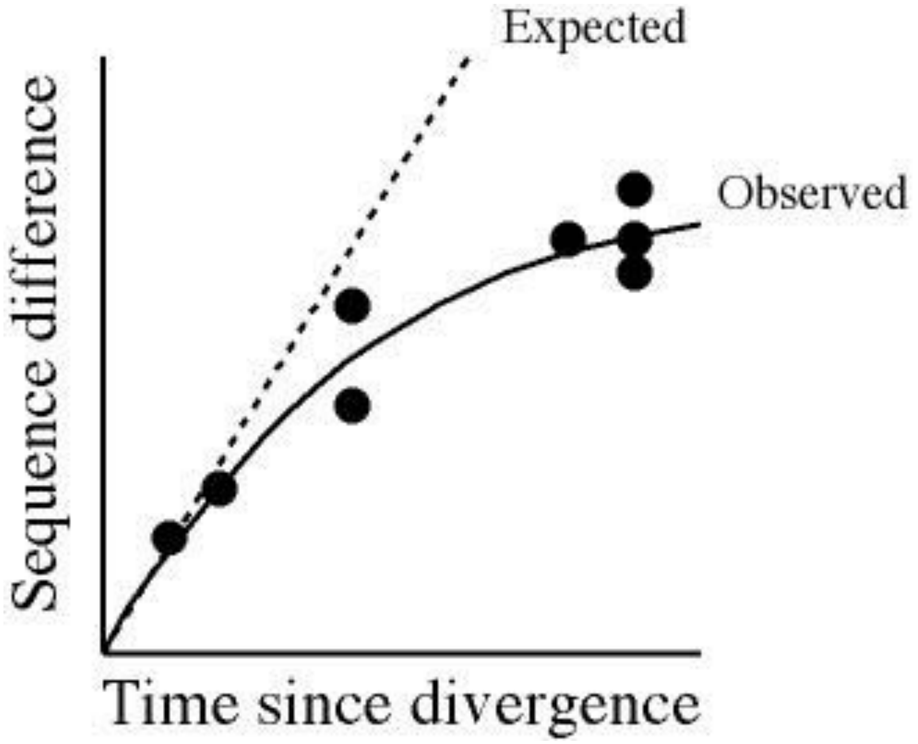
- Why are the transitions more frequent?
  - Mechanistic explanation – mutagenesis and repair mechanisms
  - Transitions are the result of:
    - tautomeric shift of bases
    - deamination (e.g. oxidative)
  - Transitions cause less distortion of the double helix structure
    - less likely to be detected and repaired by the MMR system

Modeling DNA evolution

- Ancestral sequence usually not available
- The number of mutations has to be inferred from differences between present sequences
- Requires correcting for multiple hits, particularly for more distant sequences

Calculating distance - multiple hit problem





Modelling sequence evolution

- Markov models – the state of generation n +1 depends only on the state of generation n and the rule set (character substitution probability matrix)
- There are many models with varying complexity
- Parameters may include:
  - multiple hits (Poisson correction)
  - different substitution probabilities for various mutations
  - different substitution probabilities for various ositions in sequence
  - nucleotide frequencies

DNA evolution -the simplest model (Jukes-Cantor)

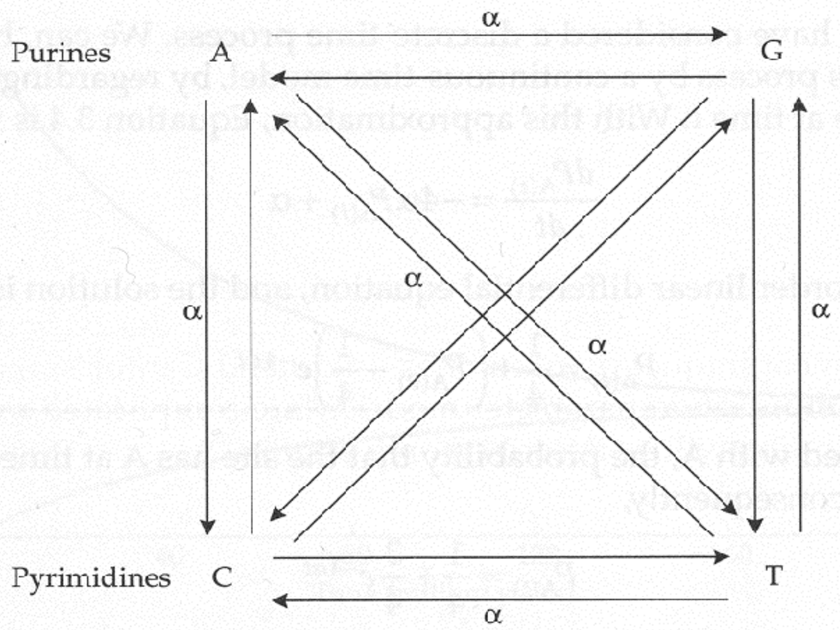


FIGURE 3.1 **One-parameter model of nucleotide substitution. The rate of substitution in each direction is α.**

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

$$D_{JC} = -\frac{3}{4}\ln(1-\frac{4}{3}D)$$

Other models

- Kimura (K80, two-parameter) - different probabilities for transitions and transversions

- Felsenstein (F81), Hasegawa-Kishino-Yano (HKY85) - different nucleotide frequencies (F81) + different probabilities for transitions and transversions (HKY85)

- GTR (General Time Reversible, Tavare ‘86)

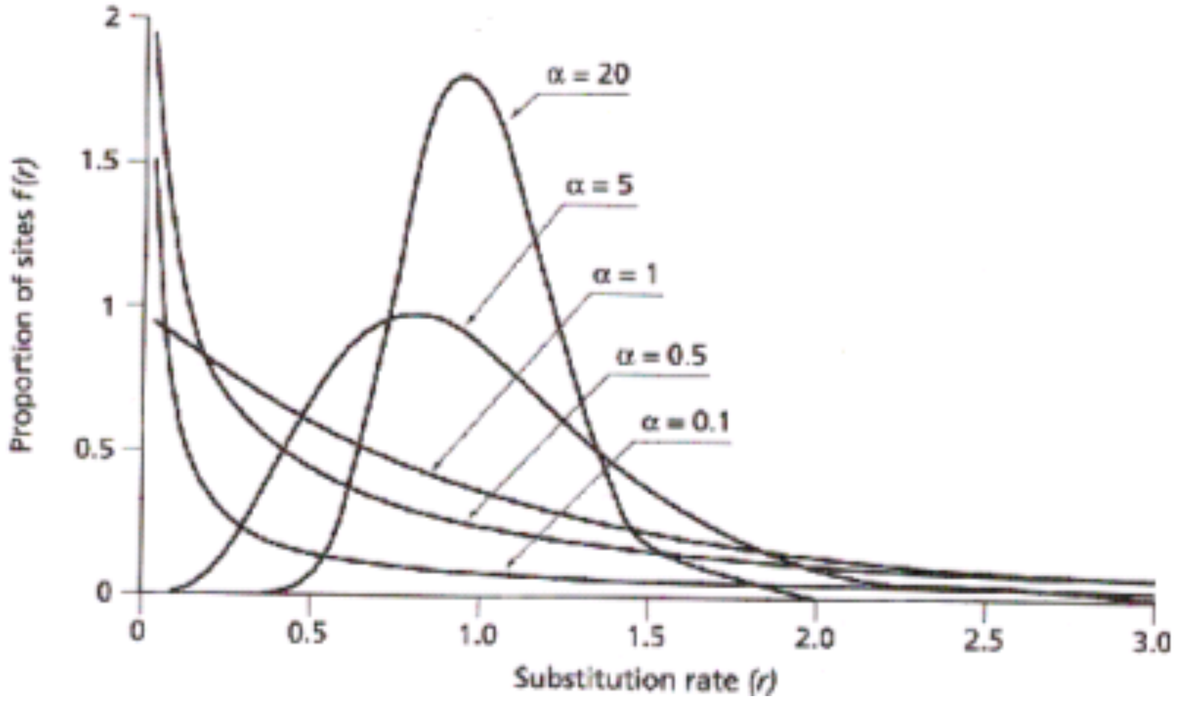
The GTR model

- Different probabilities for each substitution (but symmetrical, e.g. A→T = T→A) - 6 parameters

- Different nucleotide frequencies - 4 parameters

The gamma distribution

- In simple models each position in sequence has the same substitution probability - unrealistic
- Different classes of substitution probabilities are modelled using the gamma distribution



On the level of the genetic code

- A mutation can:
  - change the codon to another codon encoding the same aa
    - synonymous
  - change the codon to a codon encoding a different aa
    - nonsynonymous
  - change the codon to a STOP codon
    - nonsens
  - cause a frameshift
  - change gene expression

Mutations and the natural selection

- We do not observe mutations
  - we observe differences between populations (species)
  - or intra-population variation (polymorphism)
- Alleles formed by mutation are subject to selection
- Allele frequencies can be changed by genetic drift
- We observe mutations that are fixed entirely or partially (polymorphisms) in the gene pool

The fundamental question of molecular evolution

- What is the role of genetic drift and natural selection in shaping sequence variation?
  - intrapopulation (polymorphisms)
  - interspecific

- The question concerns quantitative differences!
  - There is no doubt that evolutionary adaptations are a result of natural selection!
  - But, how many of the differences we observe are adaptive?
  - And which ones?

Selection or drift?

- Selectionism
  - the majority of fixed mutations were positively selected
  - most polymorphisms are maintained by selection
    - balancing selection, overdominance, frequency-based selection

- Neutralism (Kimura, 1968)
  - the majority of fixed mutations were fixed by genetic drift (by chance)
  - the majority of polymorphisms are the result of drift
  - positively selected mutations are rare and are not significant for the quantitative analysis of molecular variation

Mutations and natural selection

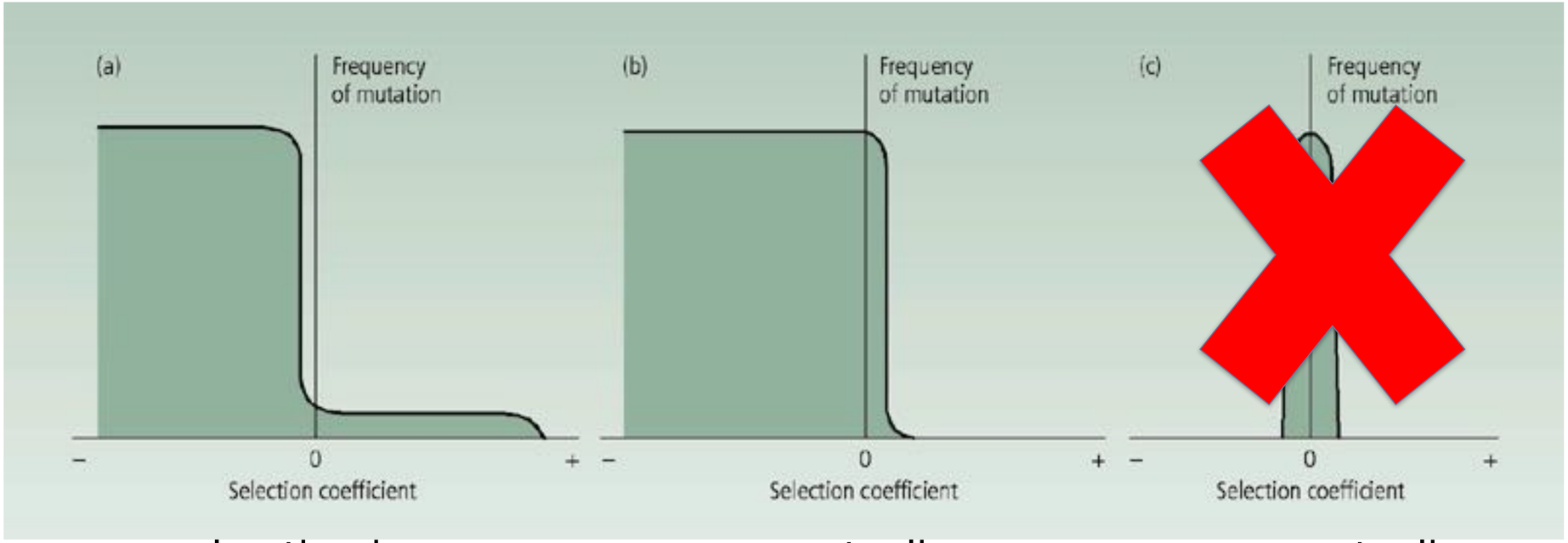
- deleterious
  - $s < 0$
  - eliminated by selection (purifying/negative)
- neutral
  - $s \approx 0$ (more precisely, $s \leq 1/4N$)
  - can be fixed or lost by genetic drift
- beneficial
  - $s > 0$
  - fixed by positive selection (influenced by drift for small s)

Selectionism vs neutralism

- Selectionism:
  - most mutations are deleterious
  - most fixed mutations are beneficial
  - neutral mutations are rare (not more frequent than beneficial)

- Neutralism
  - most mutations are deleterious or neutral
  - most fixed mutations are neutral
  - beneficial mutations are rare (much less frequent than neutral), but still important for adaptation

Selectionism vs neutralism



selectionism neutralism pan-neutralism

Neutralism is not pan-neutralism, nobody denies the selective importance of mutations!

Neutral theory - indications

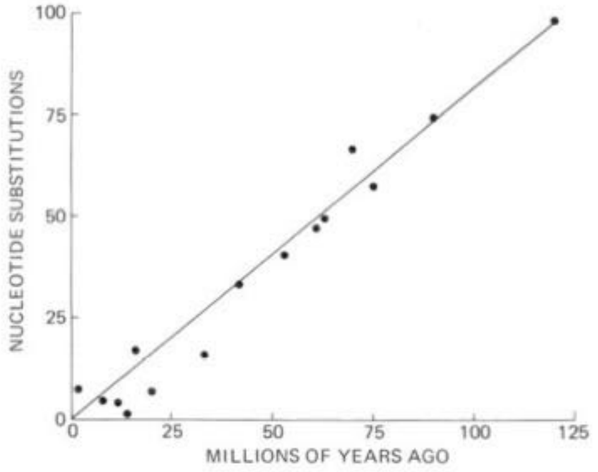
- The rate of substitution and the degree of polymorphism are too high to be explained by selection alone

- Constant rate of molecular evolution (molecular clock)

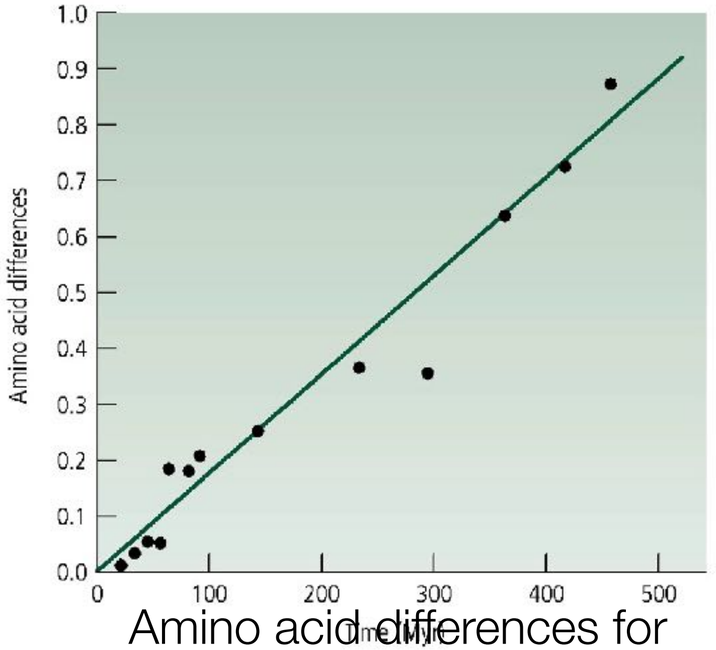
- Less important sequences (pseudogenes, less important fragments of proteins) change faster than key functional sequences

The constant rate of molecular evolution

- Many sequences evolve at a constant rate
- The rate varies for different sequences, but for the same sequence remains constant between lineages



Pairwise nucleotide differences among 17 mammals from 7 proteins, plotted against date of divergence as estimated from fossil record



Amino acid differences for vertebrate globins

- Molecular clock

Drift and the rate of evolution

- Genetic drift is a random process, but its rate is equal over long time
- Depends only on mutation rate (one fixed change per 1/μ generations)
- For selection a constant rate only if the environment changes at a constant rate (unlikely)
- The rate of adaptive changes is not constant

Molecular clock - the generation problem

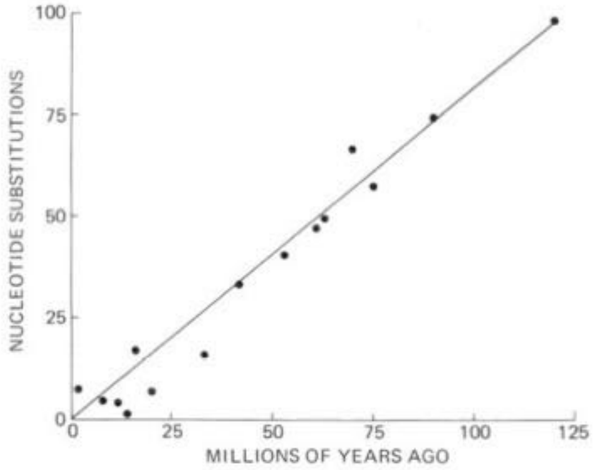
- In the neutral model the fixation rate is:

$$2N\mu\frac{1}{2N}=\mu$$

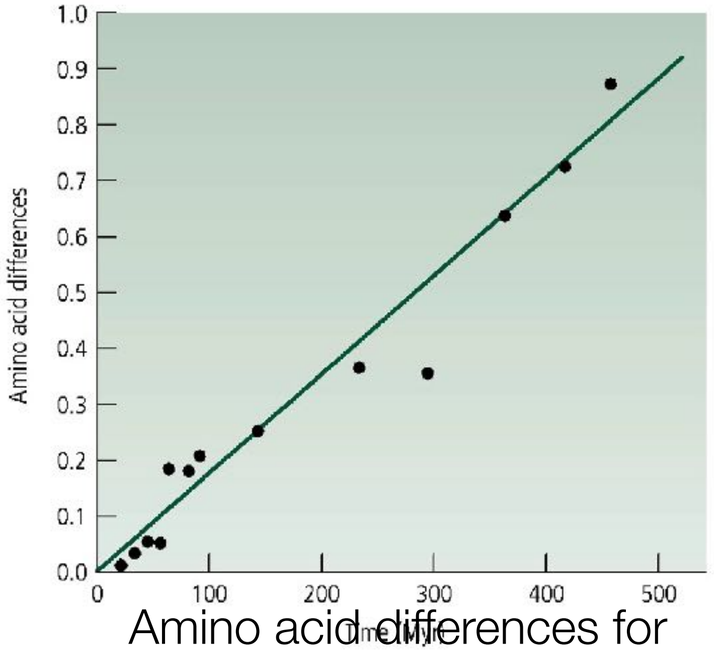
- Should be constant per generation
- Different organisms have different generation times
- The rate of evolution should not be constant over time

- But that’s what was observed

The constant rate of molecular evolution



Pairwise nucleotide differences among 17 mammals from 7 proteins, plotted against date of divergence as estimated from fossil record

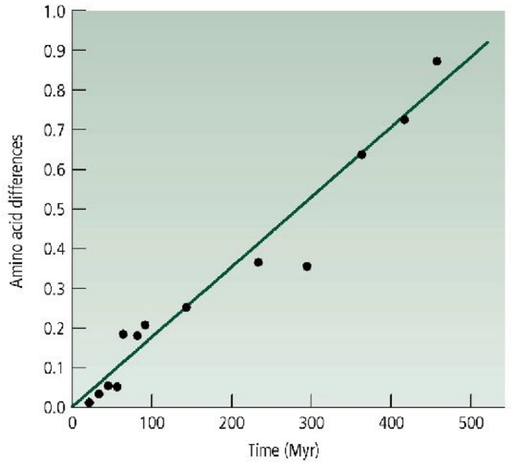


Amino acid differences for vertebrate globins

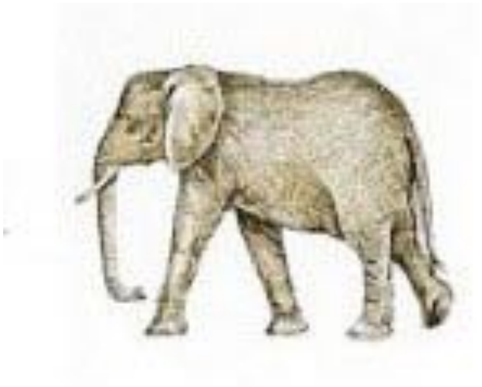
Molecular clock seems to work in real time

The generation problem

- Generation time varies among different organisms



~0.03 generations/year



~3 generations/year

- Why does the evolution rate remain constant?

The near-neutral model

- The original neutral model concerns purey neutral changes (s = 0), which are rare
- Mutations behave like neutral if:

$$|s| \leq \frac{1}{4N_e}$$

- Mutations with a small selection coefficient s will behave like neutral mutations in small populations
- in large populations they will be subject to selection

Near-neutral model

- There is a negative correlation between generation time and population size

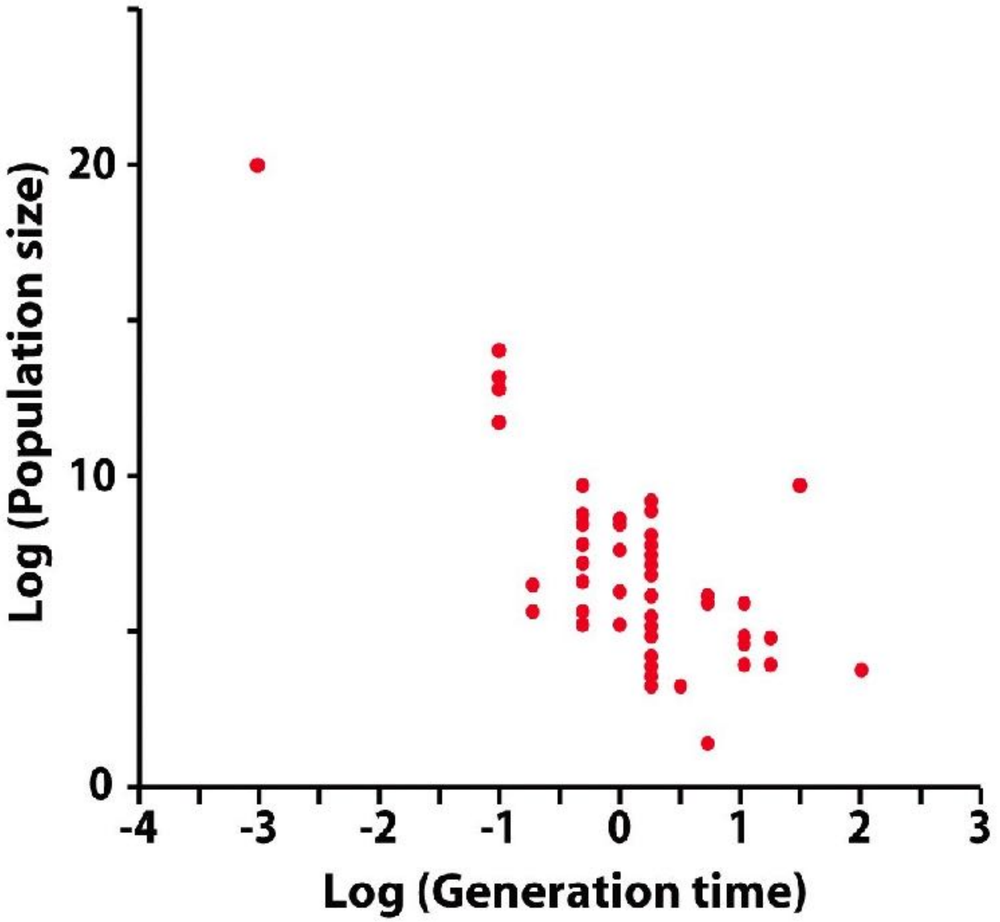
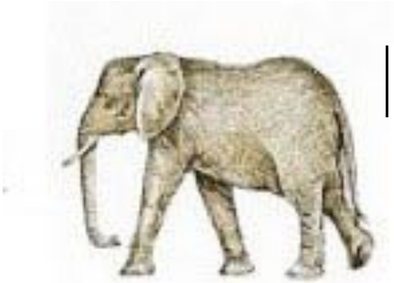


Figure 7-22a Evolutionary Analysis, 4/e

The near neutral model

~0,03 generations/year



$$|s| \leq \frac{1}{4N_e}$$

~3 generations/year

Long generation time	Short generation time
Less mutations per year	More mutations per year
Small population (small Ne)	Large population (large Ne)
More mutations behave like neutral and are fixed by drift	More mutations are subject to selection

Generation time and population size cancel each other, resulting in a constant rate (Ohta & Kimura, 1971).

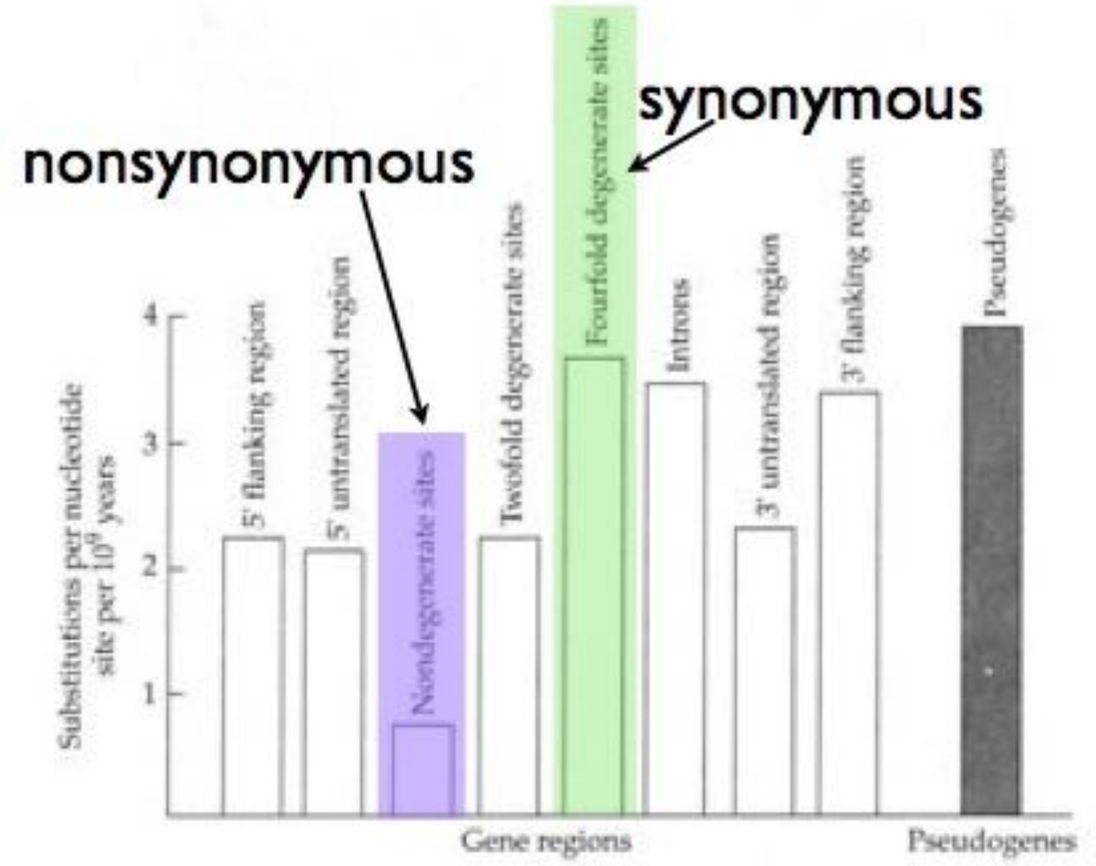
The molecular clock

- Rate constant over time for proteins and nonsynonymous substitutions

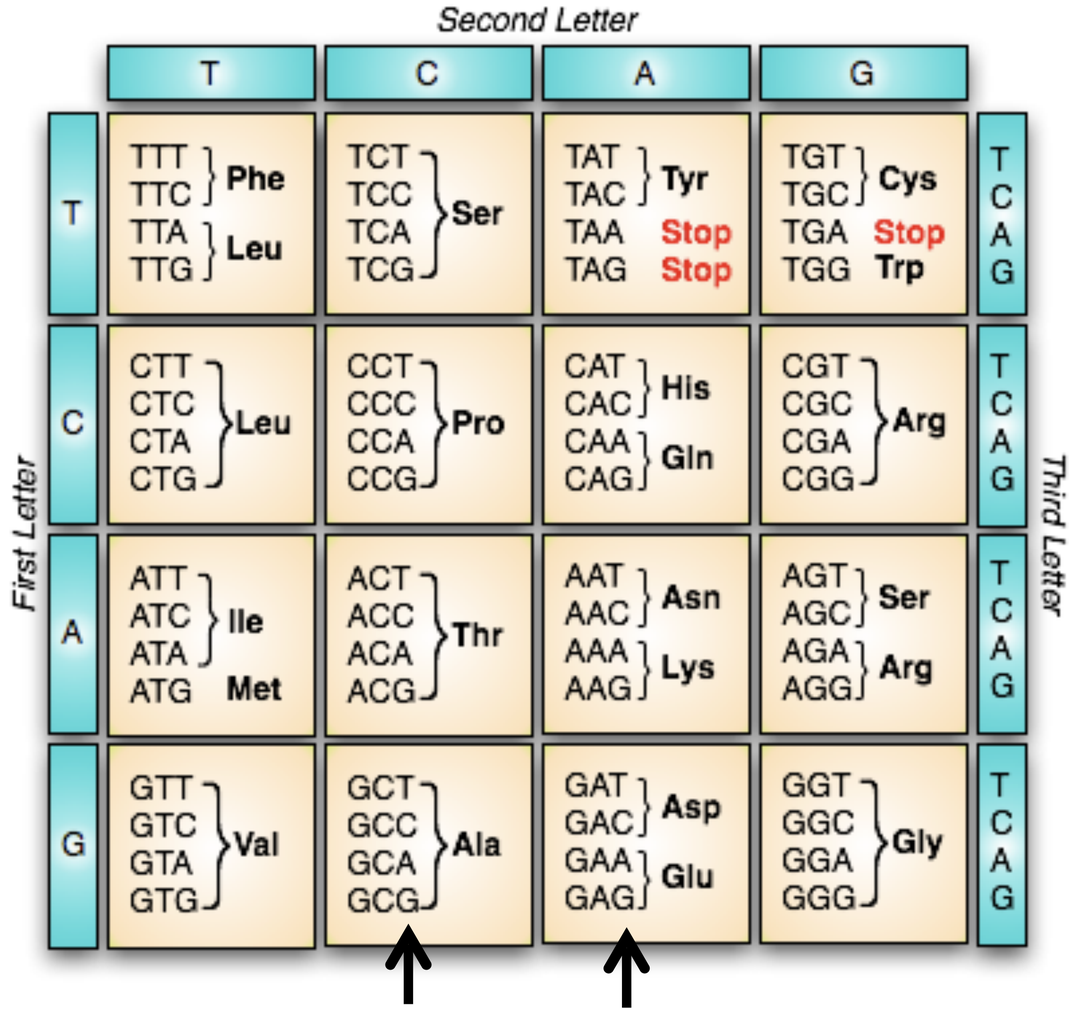
- In DNA,
  - for synonymous mutations
  - pseudogenes
  - some noncoding sequences
- the rate depends on the generation time

Eveolutionary rate and function

Less important sequences (pseudogenes, less important fragments of proteins) change faster than key functional sequences



Redundancy in the code

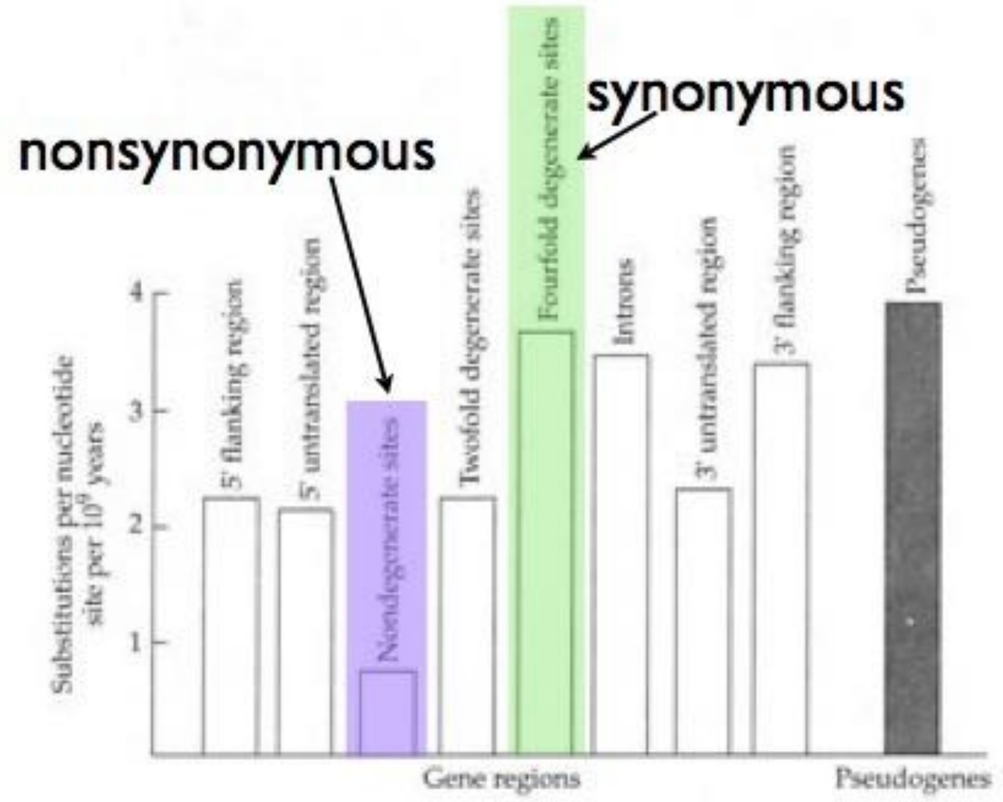


4-fold degenerate site

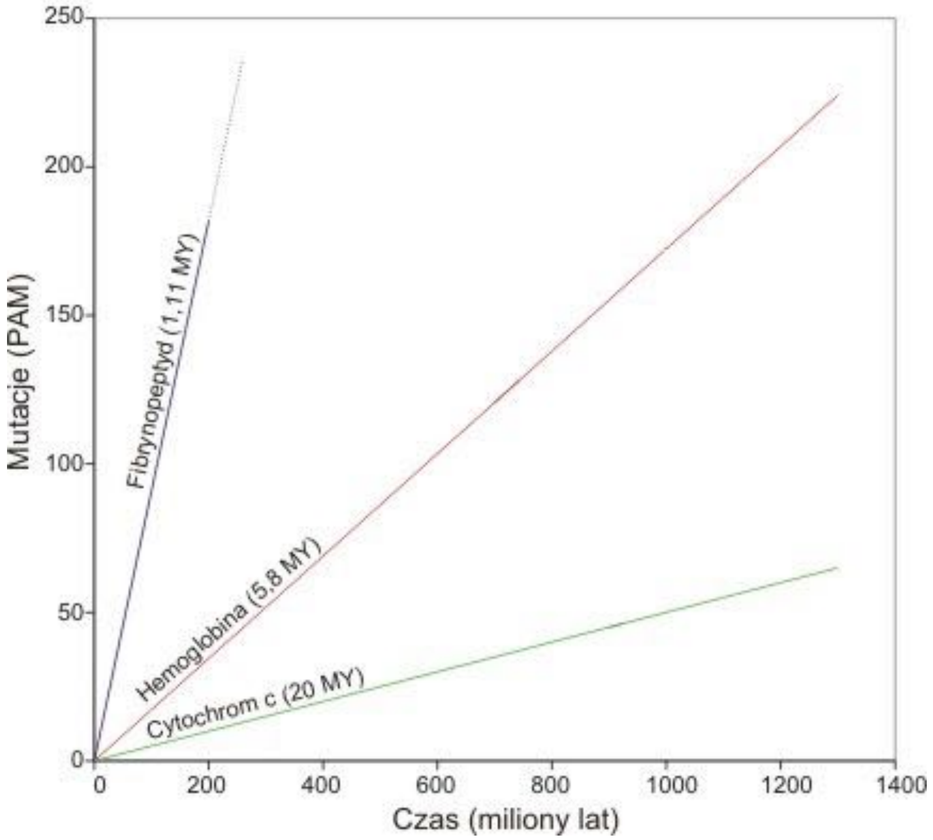
2-fold degenerate site

Eveolutionary rate and function

Less important sequences (pseudogenes, less important fragments of proteins) change faster than key functional sequences



Evolutionary rate



Evolutionary rate

- Negative (purifying selection) is the main factor influencing the rate of evolution
  - in more important sequences a mutation is more likely to be deleterious (and eliminated by selection)
  - in more important sequences a mutation is more likely to be neutral (and may be fixed by drift)
  - mutations that don’t influence function will be neutral
    - pseudogenes
    - noncoding regions?
    - synonymous substitutions?

Neutralism - the current status

- A foundation that explains many observations
  - high DNA and protein polymorphism
  - molecular clock
    - many deviations, there is no global clock, local clocks can be found
  - slower evolution of more important sequences

- It is a null hypothesis for testing for the positive selection at the molecular level!

Neutralism - the current status

- verified by sequencing, currently on the genomic scale
  - Kimura: 1968 – before DNA sequencing was invented!

Neutralism - the current status

- Smith & Eyre-Walker 2002 – 45% amino acid substitutions in Drosophila sp. fixed by selection

- Andolfatto 2005 – between *D. melanogaster* and *D. simulans* postive selection responsible for:
  - 20% DNA substitutions in introns and intragenic sequences
  - 60% DNA substitutions in UTR sequences

Neutralism - the current status

- The main achievement of the neutral theory is the development of a mathematical frameworkto sudy the effects of selection and drift

- Allowed to develop methods of testing for positive selection (using the neutral model as a null hypothesis)

- A significant portion of the genome evolves according to the neutral model
  - a local molecular clock can usually be found

The ENCODE dispute

- ENCODE - Encyclopedia of DNA Elements
- Many noncoding regions are transcribed
  - are 80% of genome functional
  - is there any “junk DNA”?
- **If there’s no selection, there’s no function!**
- Traces of selection: 2-15% of genome

ARTICLE

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many

Genome Biology and Evolution

GBE

On the Immortality of Television Sets: “Function” in the Human Genome According to the Evolution-Free Gospel of ENCODE

Dan Graur[1,*], Yichen Zheng[1], Nicholas Price[1], Ricardo B.R. Azevedo[1], Rebecca A. Zufall[1], and Eran Elhaik[2]

[1]Department of Biology and Biochemistry, University of Houston

[2]Department of Mental Health, Johns Hopkins University Bloomberg School of Public Health

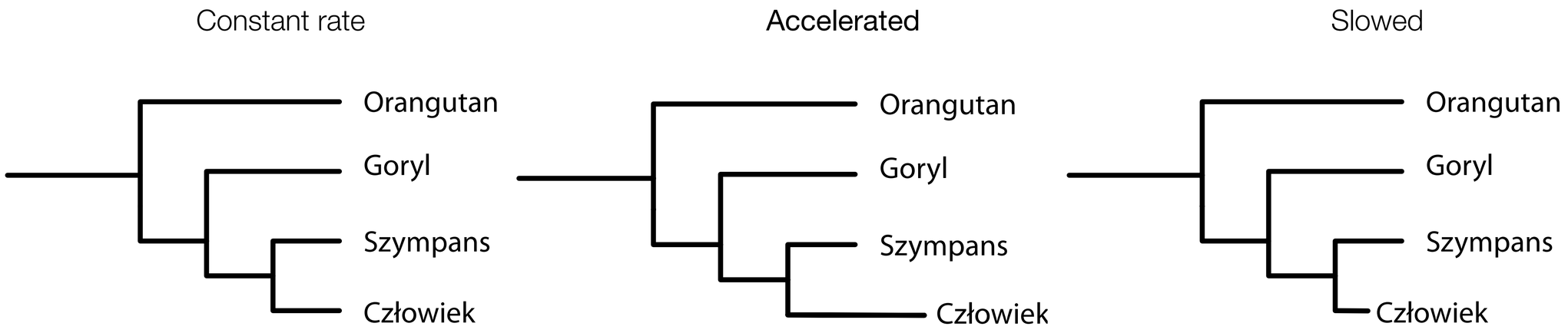
*Corresponding author: E-mail: dgraur@uh.edu.

Accepted: February 16, 2013

Testing selection

Poszukiwanie śladów działania doboru

- Most sequences evolve in a constant, clocklike fashion
- Deviation from the constant rate in a given lineage – lineage-specific selection



Badanie doboru

- Assumption: synonymous mutations are neutral
- Ka (dN) – number of nonsynonymous changes per number of nonsynonymous sites
- Ks (dS) – number of synonymous changes per number of synonymous sites
- Ka/Ks (ω) ratio is the measure of selection

ω=1 – purely neutral
ω<1 – negative (purifying) selection
ω>1 – positive selection

Testing for selection

- The ω rate is rarely larger than 1 globally for the entire gene (exceptions, e.g. MHC genes)
- Average ω between primates and rodents is 0.2, between human and chimp: 0.4
- Deviations from average ω for a given gene in a given lineage indicate selection
- In a gene there can be sites with different ω, indicating selection acting on particular regions of the sequence

Testing for selection II

- Comparing synonymous and nonsynonymous rates for intraspecific and interspecific comparisons

The McDonald-Kreitman test

- Comparing synonymous and nonsynonymous rates for intraspecific and interspecific comparisons

- In a neutral model both rates should be equal

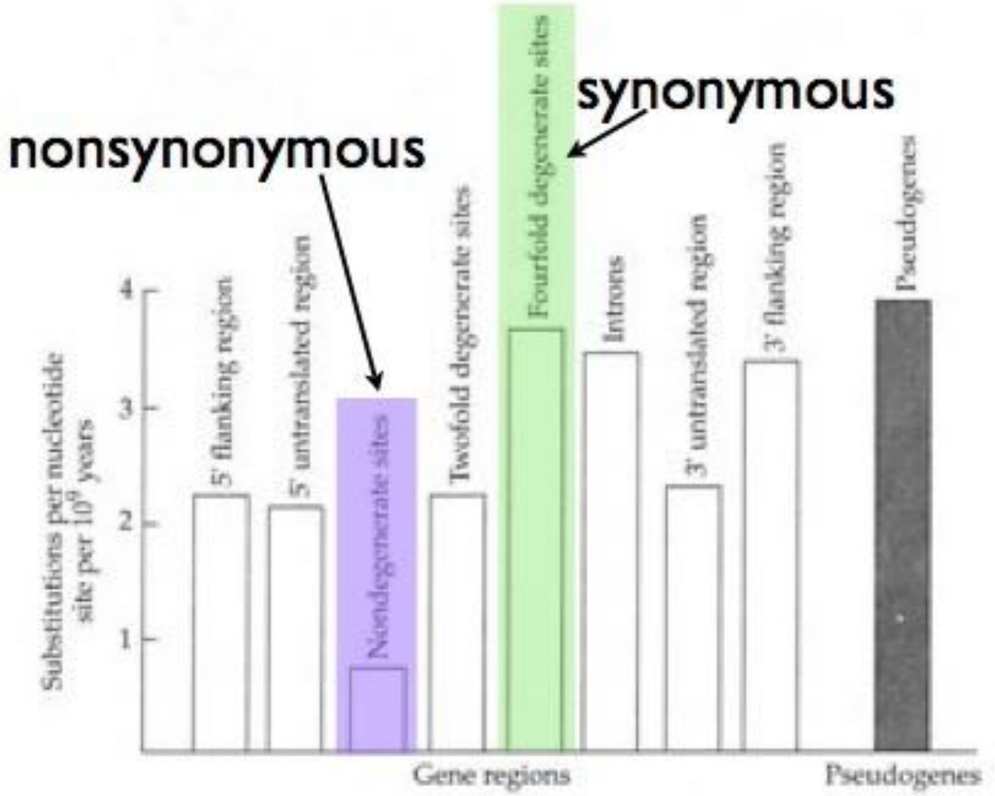
Example: the ADH gene in 3 Drosophila species

	synonymous	nonsynonymous	rate
intraspecific	42	2	~0.05
interspecific	17	7	~0.41

Conclusion: nonsynonymous changes favoured during speciation - not neutral

Are the synonymous changes truly neutral?

- There is some selection on 4-fold degenerate sites



Are the synonymous changes truly neutral?

- Synonymous codons are not equivalent - rare and frequent codons

- Changing a frequent codon to a rare synonymous codon can influence expression levels and kinetics

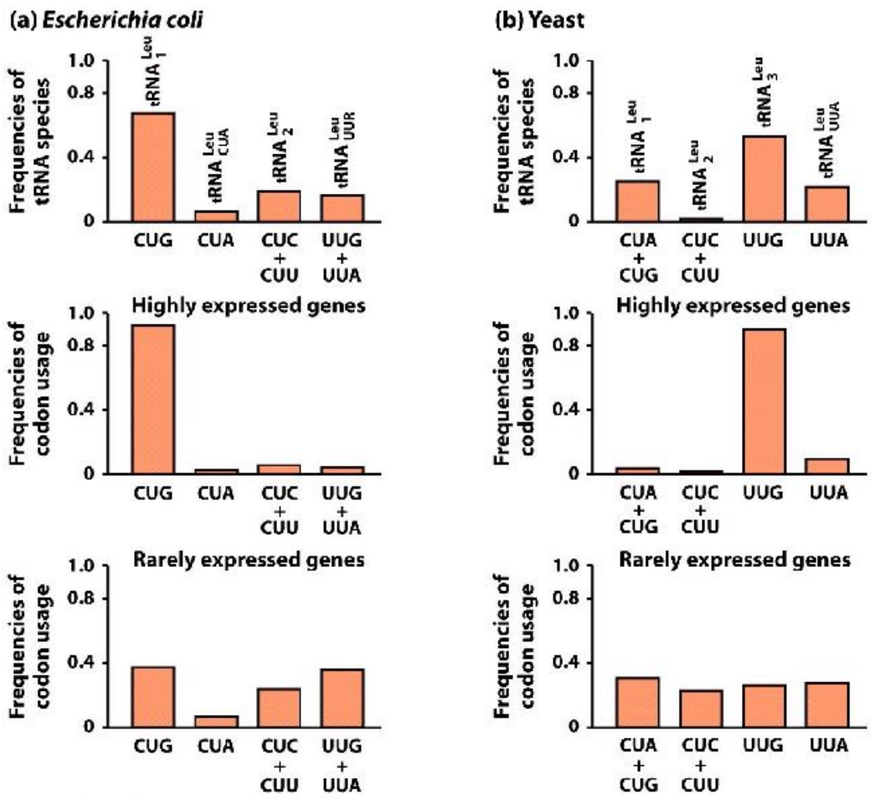


Figure 7-24 Evolutionary Analysis, 4/e

Are the synonymous changes truly neutral?

RESEARCH ARTICLE

AMERICAN JOURNAL OF medical genetics PART A

A Synonymous Mutation in *TCOF1* Causes Treacher Collins Syndrome Due to Mis-Splicing of a Constitutive Exon

D. Macaya,[1] S.H. Katsanis,[1] T.W. Hefferon,[2] S. Audlin,[1] N.J. Mendelsohn,[3] J. Roggenbuck,[3] and G.R. Cutting[1]*

[1]DNA Diagnostic Laboratory, Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland

[2]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland

[3]Children's Hospitals & Clinics of Minnesota, Minneapolis, Minnesota

Received 26 September 2008; Accepted 25 February 2009

Are the synonymous changes truly neutral?

A "Silent" Polymorphism in the *MDR*1 Gene Changes Substrate Specificity

Chava Kimchi-Sarfaty,*† Jung Mi Oh,†‡ In-Wha Kim, Zuben E. Sauna,
Anna Maria Calcagno, Suresh V. Ambudkar, Michael M. Gottesman†

Synonymous single-nucleotide polymorphisms (SNPs) do not produce altered coding sequences, and therefore they are not expected to change the function of the protein in which they occur. We report that a synonymous SNP in the *Multidrug Resistance* 1 (*MDR*1) gene, part of a haplotype previously linked to altered function of the *MDR*1 gene product P-glycoprotein (P-gp), nonetheless results in P-gp with altered drug and inhibitor interactions. Similar mRNA and protein levels, but altered conformations, were found for wild-type and polymorphic P-gp. We hypothesize that the presence of a rare codon, marked by the synonymous polymorphism, affects the timing of cotranslational folding and insertion of P-gp into the membrane, thereby altering the structure of substrate and inhibitor interaction sites.

The *MDR*1 gene product, the adenosine triphosphate (ATP)–binding cassette (ABC) transporter ABCB1 or P-gp, is an ATP-driven efflux pump contributing to the pharmacokinetics of drugs that are P-gp substrates and to the multidrug resistance of cancer cells (*1*, *2*). To date, more than 50 single-nucleotide polymorphisms (SNPs) have been reported for *MDR*1 (www.ncbi.nlm.nih.gov/SNP/GeneGt.cgi?geneID=5243). One of these, a synonymous SNP in exon 26 (C3435T), was sometimes found to be associated with altered P-gp activity (*3*–*6*) and, when it appears in a haplotype, with reduced functionality (*7*). This association may be explained in different ways. Perhaps it is because C3435T is in linkage disequilibrium with other common functional nonsynonymous polymorphisms such as G2677T. In fact, the C1236T (a synonymous SNP), G2677T, and C3435T polymorphisms are part of a common haplotype (*8*, *9*). Another possible explanation is that allele-specific differences in mRNA folding could influence splicing, processing, or translational control and regulation (*10*, *11*). A third possibility is that the effect of the C3435T polymorphism on the levels of cell surface P-gp activity or its function is rather modest or drug-specific. Finally, numerous environmental factors are known to affect the expression and phenotypic activity of P-gp (*12*).

To determine whether the C3435T polymorphism actually does affect P-gp activity, we expressed wild-type and polymorphic P-gps in HeLa cells with the use of a transient expression system (*13*). The same experiments were carried out on BSC-1 (epithelial cells of African green monkey kidney origin), Vero-76 (monkey kidney cells), and 12E1 (CEM human cells) cell lines (*14*), with similar results, indicating that this phenomenon is not specific to HeLa cells.

Laboratory of Cell Biology, Center for Cancer Research, National Cancer Institute, Bethesda, MD 20892, USA.

*Present address: Center for Biologics Evaluation and Research, Food and Drug Administration, 29 Lincoln Drive, Room 316, Bethesda, MD 20892, USA.

†To whom correspondence should be addressed. E-mail: mgottesman@nih.gov (M.M.G.); jmoh@snu.ac.kr (J.M.O.); kimchi@cber.fda.gov (C.K.-S.)

‡Present address: College of Pharmacy, Seoul National University, Seoul 151-742, South Korea.