Mutation Rates and Sequence Changes

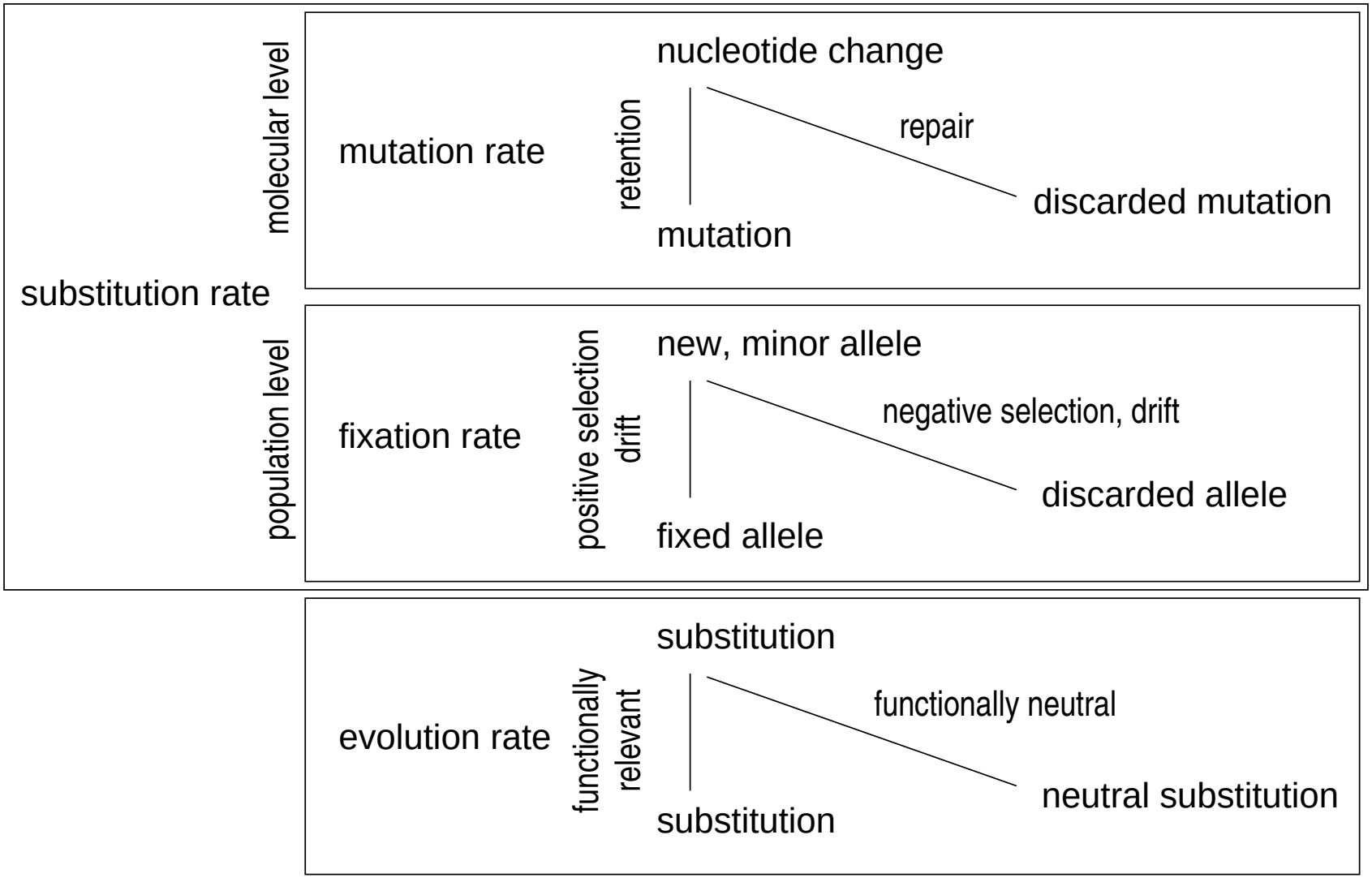
part of “Fortgeschrittene Methoden in der Bioinformatik”

Sonja Prohaska

Computational EvoDevo
University Leipzig

Leipzig, WS 2011/12

From Molecular to Population Genetics



Nucleotide Exchanges

transition: exchange purine for purine (C $\leftrightarrow$ T) or pyrimidine for pyrimidine (A $\leftrightarrow$ G)
transversion: exchange purine for pyrimidine or pyrimidine for purine (C | T $\leftrightarrow$ A | G)
synonymous substitution: nucleotide changes that are functionally neutral
nonsynonymous substitution: nucleotide changes that change the function

Estimating Mutation Rates

- take two species that diverged a time T ago (i.e. had a common ancestor a time T ago)
- select regions that
  - are 1:1 orthologs of each other (i.e. have a common ancestral sequence in the common ancestor and were not duplicated since)
  - evolved neutrally (i.e. were not under positive or negative selection since their divergence from the common ancestor)
  - can be aligned without errors
- count the number of substitutions
- correct for reversion and multiple mutations at the same site and biases
- devide the number of nucleotide exchanges (mutations) by T

Purifying *versus* Positive Selection I

- Selection can only occure at nonsynonymous sites.
- Mutations fixed by **purifying selection**: the rate of fixation of synonymous changes is greater than the rate of fixation of nonsynonymous changes ($\omega_S < 1$).
- Mutations fixed by **positive selection**: the rate of fixation of nonsynonymous changes is greater than the rate of fixation of synonymous changes ($\omega_S > 1$).

$$\omega_S = \frac{d_N}{d_S} \tag{1}$$

ω_S ... selection ratio
d_s ... synonymous divergence per synonymous site
d_N ... nonsynonymous divergence per nonsynonymous site

Purifying *versus* Positive Selection II

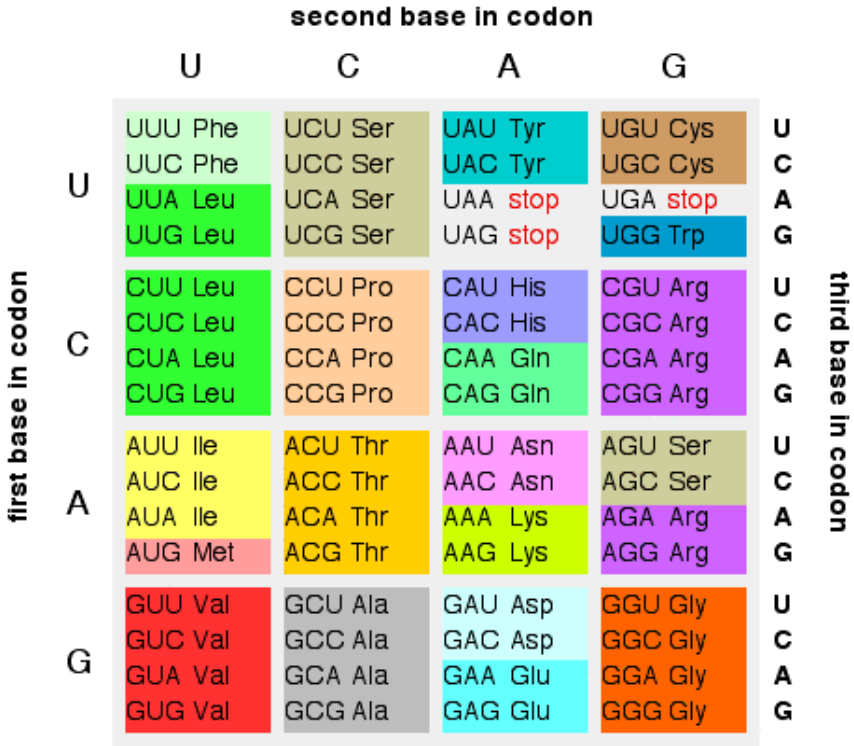
The following would be more accurate:

$$\omega = \frac{d_N/2T}{\mu_N} \tag{2}$$

The selection ratio ω is the ratio of the rate of nonsynonymous substitutions per site d_N to the rate of nonsynonymous mutations per site μ_N.

How can we estimate μ_N?

4-fold Degenerate Sites



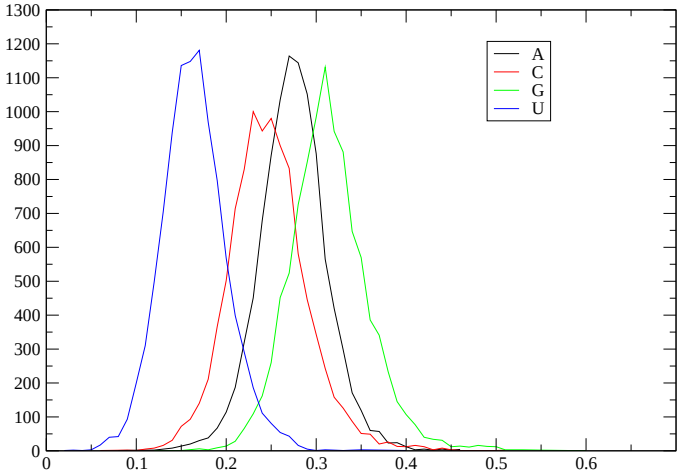
?-fold degenerate site: ? = the number of different nucleotides that can occure at the site without changing the protein sequence

CU*	Leu	GU*	Val	UC*	Ser	CC*	Pro
AC*	Thr	GC*	Ala	CG*	Arg	GG*	Gly

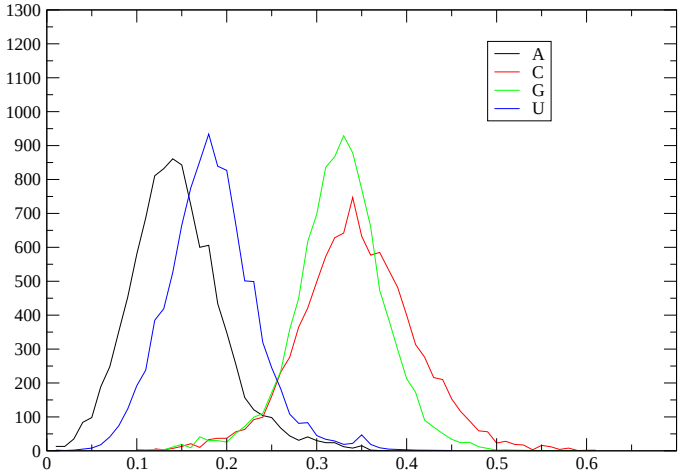
Assumption: 4-fold degenerate sites are synonymous sites.

Nucleotide Occurence at Codon Positions in *Drosophila melanogaster*

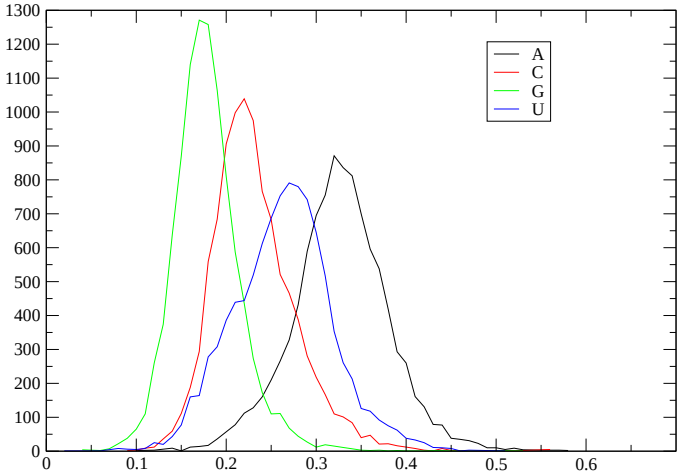
1st codon position



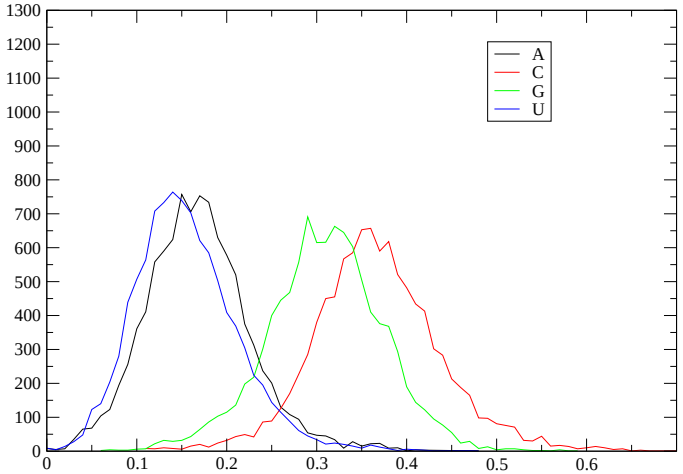
3rd codon position



2nd codon position 0-fold degenerate



part of 3rd codon position 4-fold degenerate



Why are nucleotide frequences different for different codon positions?

Potential Causes

- codon usage bias
- base composition bias
- selective constraints on other levels than the coding sequence

Estimating the Codon Usage Bias I

Relative Synonymous Codon Usage (RSCU)

$$E(X_{ij}) = \frac{\sum_j X_{ij}}{n_{ij}} \tag{3}$$

$$RSCU_{ij} = \frac{X_{ij}}{E(X_{ij})} = X_{ij}/(1/n_i \sum_{j=1}^{n} X_{ij}) \tag{4}$$

i	...	index running over the 20 amino acids
j_i	...	index running over the codons for amino acid i
n_{ij}	...	the number of different codons for amino acid i
X_{ij}	...	observed number of codon j for amino acid i

$RSCU_{ij} = 1$ usage of codon j is neither prefered nor avoided
$RSCU_{ij} > 1$ codon j is used preferentially
$RSCU_{ij} < 1$ codon j is avoided

Estimating the Base Composition Bias

Base Composition Skew (BCS)

$$BCS = \sum_{n_i \in \{ACGT\}} (n_i - E(n_i))^2 \quad (5)$$

Sum of the squared deviation of the observed nucleotide frequency from the expected nucleotide frequency $E(n_A) = E(n_T) = E(n_C) = E(n_G) = 0.25$.

Genomic Mutation Distances

$$d_{Sg} = (1 - f_g) d_{\mu g} \tag{6}$$

d_{Sg} ... synonymous distance for gene g acording to the Tamura-Nei model

f_g ... fraction of mutations underestimated due to biases

$d_{\mu g}$ ... mutation distance for gene g

$$f_g = \eta BCS_g \tag{7}$$

BCS_g ... base compostion skew for gene g

η ... obtaind by divinding the absolute value of the slope of the linear regression of BSC on d_S by the y-intercept of the regression line

References

- [Filipski, 2008] Alan Filipski, Sonja J. Prohaska and Sudhir Kumar. *Molecular Signatures of Adaptive Evolution*. in "Evolutionary Genomics and Proteomics" edited by Mark Pagel; Sinauer Associates, Inc. Sunderland 2008. Chapter 11, p241-254.
- [Tamura, 2004] Koichiro Tamura, Sankar Subramanian and Sudhir Kumar. *Temporal Patterns of Fruit Fly (Drosophila) Evolution Revealed by Mutation Clock*. Mol. Biol. Evol. 2004. 21(1), p36-44.