Mutation Rates and Sequence Changes
part of “Fortgeschrittene Methoden in der Bioinformatik”

Sonja Prohaska
Professorship for Bioinformatics
University Leipzig

Leipzig, WS 2010/11
From Molecular to Population Genetics

- **Mutation Rate**
  - **Molecular Level**
    - Mutation rate
      - Retention
        - Repair
          - Discarded mutation
    - Nucleotide change

- **Fixation Rate**
  - **Population Level**
    - Positive selection
      - Drift
        - Negative selection, drift
        - Discarded allele
    - New, minor allele
      - Fixed allele

- **Evolution Rate**
  - Substitution
    - Functionally relevant
      - Functionally neutral
        - Neutral substitution
**Nucleotide Exchanges**

**transition**: exchange purine for purine (C ↔ T) or pyrimidine for pyrimidine (A ↔ G)

**transversion**: exchange purine for pyrimidine or pyrimidine for purine (C | T ↔ A | G)

**synonymous substitution**: nucleotide changes that are functionally neutral

**nonsynonymous substitution**: nucleotide changes that change the function
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

divide the number of nucleotide exchanges (mutations) by $T$
Selection can only occur 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 ($\omega_S > 1$).

\[
\omega_S = \frac{d_N}{d_S}
\]  

$\omega_S$ ... selection ratio  
$d_s$ ... synonymous divergence per synonymous site  
$d_N$ ... nonsynonymous divergence per nonsynonymous site
The following would be more accurate:

\[ \omega = \frac{d_N}{2T \mu_N} \]  

(2)

The selection ratio \( \omega \) is the ratio of the rate of nonsynonymous substitutions per site \( d_N \) to the rate of nonsynonymous mutations per site \( \mu_N \).

How can we estimate \( \mu_N \)?
4-fold Degenerate Sites

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

<table>
<thead>
<tr>
<th>first base in codon</th>
<th>second base in codon</th>
<th>third base in codon</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>U</td>
<td>U</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
</tbody>
</table>

Assumption: 4-fold degenerate sites are synonymous sites.
Nucleotide Occurrence at Codon Positions in *Drosophila melanogaster*

![Graphs showing nucleotide occurrence at codon positions in Drosophila melanogaster.](image_url)
## Why are nucleotide frequencies different for different codon positions?

### Potential Causes

- codon usage bias
- base composition bias
- selective constraints on other levels than the coding sequence
Relative Synonymous Codon Usage (RSCU)

\[ E(X_{ij}) = \frac{\sum_j X_{ij}}{n_{ij}} \]  \hspace{1cm} (3)

\[ RSCU_{ij} = \frac{X_{ij}}{E(X_{ij})} = \frac{X_{ij}}{X_{ij}/(1/n_i \sum_{j=1}^n X_{ij})} \]  \hspace{1cm} (4)

- \( i \) \( \ldots \) index running over the 20 amino acids
- \( j_i \) \( \ldots \) index running over the codons for amino acid \( i \)
- \( n_{ij} \) \( \ldots \) the number of different codons for amino acid \( i \)
- \( X_{ij} \) \( \ldots \) observed number of codon \( j \) for amino acid \( i \)

- \( RSCU_{ij} = 1 \) usage of codon \( j \) is neither preferred nor avoided
- \( RSCU_{ij} > 1 \) codon \( j \) is used preferentially
- \( RSCU_{ij} < 1 \) codon \( j \) is avoided
Base Composition Skew (BCS)

\[ BCS = \sum_{n_i \in \{ACGT\}} (n_i - E(n_i))^2 \]  

(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

\[ d_{Sg} = (1 - f_g)d_{\mu g} \]  \hspace{1cm} (6)

- \( d_{Sg} \) … synonymous distance for gene \( g \) according to the Tamura-Nei model
- \( f_g \) … fraction of mutations underestimated due to biases
- \( d_{\mu g} \) … mutation distance for gene \( g \)

\[ f_g = \eta BCS_g \]  \hspace{1cm} (7)

- \( BCS_g \) … base composition skew for gene \( g \)
- \( \eta \) … obtained by dividing the absolute value of the slope of the linear regression of \( BSC \) on \( d_S \) by the \( y \)-intercept of the regression line