Rates of nucleotide substitution in coding regions

So far we have talked about substitution in protein coding genes

Rates vary among genes (examples on Table 4.1)

Interleukin (fairly fast)

Ubiquitin (slow)

1 - 3 substitutions per site per billion years

Essentially zero!

Rates of nucleotide substitution in noncoding regions

What about rates in non-coding sequences?

Mostly data are from 5' and 3' untranslated regions of genes and also from pseudogenes (functionally defunct relatives of known genes).

Although rates vary, for mammals we’re still in the range of 1 - 10 substitutions per site per billion years.

Rates of nucleotide substitution in noncoding regions

Substitution rates in the genome tend to rank:

- Pseudogenes
- 4 fold degenerate sites
- Introns
- 3' and 5' untranslated regions
- 2 fold degenerate sites
- Nondegenerate sites

Causes of variation in substitution rates

Probability of substitution depends on:

1) rate of mutation

2) probability of mutation fixation

which in turn depends on:

- strength of selection
  - drift (population size)
  - draft (linkage)

selective (or functional) constraint:

intensity of purifying selection
Selective constraint

Differences in mutation rates across the genome are not enough to account for approximately 1000-fold range in nonsynonymous substitution rates across genes.

Selective constraint due to purifying selection likely to be very important

Similarity profiles

We have used measures of nucleotide diversity to infer diversifying (or balancing) selection, but the flip side is using nucleotide identity to infer selective constraint.

More functionally important genes (or parts of genes) will be more constrained and will evolve more slowly.

We can use “similarity profiles” to visualize this (Fig 4.4)

Similarity profiles

0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9

Window Midpoint Location (bp)

Nonsynonomous Substitutions (K_A)

Reptiles Mammals

Similarity profiles

Signal peptide and C chain are excised and do not participate in hormonal activity of insulin.

They are likely less constrained and have a higher nonsynonymous substitution rate than the A and B chains.

Summary variation in substitution rates

So we see variation in substitution rates among:

- types of nucleotide sites (e.g. syn vs nonsyn)
- regions within genes
- genes
- organisms

Similarity profiles

G+L warn against circularity, but similarity profiles are often “read backwards” and rates of substitution are used to infer the level of constraint of sequences.
Summary variation in substitution rates

So we see variation in substitution rates among:

- types of nucleotide sites (e.g. syn vs nonsyn)
- regions within genes
- genes
- organisms

Example: sequences in RNA viruses may evolve up to 1 million times faster than those in DNA genomes

Method for estimating rate of substitution in viruses on pg 161

RNA virus example

High mutation rates in HIV may be due to:

- errors in reverse transcription
- very little purifying selection

Amino acid replacement

We have been talking about nucleotide substitution, but analogous patterns are observed for amino acid replacement.

Replacement by similar a.a. --> conservative replacement

Replacement by dissimilar a.a. --> radical replacement

Amino acid replacement

Conservative replacements are more common than radical replacements (similar to how syn subs are more common than nonsyn subs)

Stability index: how similar an a.a. is to its single step mutational derivatives

Amino acid exchangeability circle (Fig 4.12)

Protein evolution

Celeste will be giving guest lecture about protein evolution

There are some elements of proteins that are generally conserved in evolution:

- volume
- density
- hydrophobicity
- polarity

Functional constraints in proteins obviously effect patterns of nucleotide substitution and amino acid replacement.
Nucleotide composition

If mutation rates among all bases are equal, we would expect A+T and G+C frequencies in the genome to be 50% each.

But this is often not observed.

Commonly observed mutations are a subset of all possible base-substitution mutations.

In fact the G-C content of genomes varies among regions and among organisms (we will talk about this more later in the semester).

Example: figure 4.5 pseudogenes - transitions occur more frequently than transversions (e.g., C-->T, G-->A).

<table>
<thead>
<tr>
<th>Base</th>
<th>Proportion of mutation TO given base</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>27.9</td>
</tr>
<tr>
<td>T</td>
<td>28.5</td>
</tr>
<tr>
<td>C</td>
<td>23.2</td>
</tr>
<tr>
<td>G</td>
<td>20.5</td>
</tr>
</tbody>
</table>

Also, from a metabolic perspective, Cs and Gs are most costly to produce than As and Ts.

However G:C pairs have 3 hydrogen bonds and A:T pairs have 2 hydrogen bonds so genomic stability should increase with higher G-C content.

Codon usage bias

Even synonymous codons are not used with equal frequency.
**Codon usage bias**

**Relative Synonymous Codon Usage**

\[
RSCU = \frac{X_i}{\frac{1}{n} \sum_{i=1}^{n} X_i}
\]

*observed occurrences of codon*  
*expected occurrences of codon under equal usage*

If RSCU is NOT equal to 1, there is codon usage bias

There are other methods for calculating preferred codons

**Codon usage bias**

A subset of hypotheses for codon usage bias:

- **Translational efficiency**: use codons that are recognized by the most abundant tRNAs
- **Replicational/translational stability**: codons with mononucleotide runs may be vulnerable to slippage so selectively eliminated

---

**Male vs female mutational input**

**Spermatogenesis vs oogenesis**

Many more germ cell divisions in males than females - possibly greater [mutational input](#) by males

---

**Reading for next class**

Molecular clock section in chap 4

Bromham and Penny 2003 paper

---

**Male vs female mutational input**

**Studies of zinc finger genes**

In birds females are the heterogametic sex *(WZ instead of WW)*

Rates of evolution in males still higher than in females

---

**Male vs female mutational input**

Studies of zinc finger genes

In mammals males are the heterogametic sex.

Zfy is Y-linked and Zfx if X-linked

Zfy evolving 6x faster than Zfx, corresponding to approximately 6x greater number cell divisions