part 3: analysis of natural selection pressure
types of codon models

\[ Q_{ij} = \begin{cases} 
0 & \text{if } i \text{ and } j \text{ differ by } > 1 \\
\pi_j & \text{for synonymous tv.} \\
\kappa \pi_j & \text{for synonymous ts.} \\
\omega \pi_j & \text{for non-synonymous tv.} \\
\omega \kappa \pi_j & \text{for non-synonymous ts.} 
\end{cases} \]

“OMEGA MODELS”

Goldman and Yang (1994)
Muse and Gaut (1994)
this codon model “M0”

“OMEGA MODELS”

\[ Q_j = \begin{cases} 
0 & \text{if } i \text{ and } j \text{ differ by } > 1 \\
\pi_j & \text{for synonymous } t_v. \\
\kappa\pi_j & \text{for synonymous } t_s. \\
\omega\pi_j & \text{for non-synonymous } t_v. \\
\omega\kappa\pi_j & \text{for non-synonymous } t_s.
\end{cases} \]

Goldman and Yang (1994)
Muse and Gaut (1994)

same \( \omega \)
for all branches

same \( \omega \)
for all sites
two basic types of models

branch models
(\( \omega \) varies among branches)

site models
(\( \omega \) varies among sites)
interpretation of a branch model

episodic adaptive evolution of a novel function with $\omega_1 > 1$
**branch models**

<table>
<thead>
<tr>
<th>variation (ω) among branches:</th>
<th>approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang, 1998</td>
<td>fixed effects</td>
</tr>
<tr>
<td>Bielawski and Yang, 2003</td>
<td>fixed effects</td>
</tr>
<tr>
<td>Seo et al. 2004</td>
<td>auto-correlated rates</td>
</tr>
<tr>
<td>Kosakovsky Pond and Frost, 2005</td>
<td>genetic algorithm</td>
</tr>
<tr>
<td>Dutheil et al. 2012</td>
<td>clustering algorithm</td>
</tr>
</tbody>
</table>

* these methods can be useful when selection pressure is strongly **episodic**
### Site Models

<table>
<thead>
<tr>
<th>Variation (ω) Among Sites</th>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang and Swanson, 2002</td>
<td>fixed effects (ML)</td>
</tr>
<tr>
<td>Bao, Gu and Bielawski, 2006</td>
<td>fixed effects (ML)</td>
</tr>
<tr>
<td>Massingham and Goldman, 2005</td>
<td>site wise (LRT)</td>
</tr>
<tr>
<td>Kosakovsky Pond and Frost, 2005</td>
<td>site wise (LRT)</td>
</tr>
<tr>
<td>Nielsen and Yang, 1998</td>
<td>mixture model (ML)</td>
</tr>
<tr>
<td>Kosakovsky Pond, Frost and Muse, 2005</td>
<td>mixture model (ML)</td>
</tr>
<tr>
<td>Huelsenbeck and Dyer, 2004; Huelsenbeck et al. 2006</td>
<td>mixture (Bayesian)</td>
</tr>
<tr>
<td>Rubenstein et al. 2011</td>
<td>mixture model (ML)</td>
</tr>
<tr>
<td>Bao, Gu, Dunn and Bielawski 2008 &amp; 2011</td>
<td>mixture (LiBaC/MBC)</td>
</tr>
<tr>
<td>Murell et al. 2013</td>
<td>mixture (Bayesian)</td>
</tr>
</tbody>
</table>

• Useful when at some sites evolve under **diversifying selection** pressure over long periods of time

• This is not a comprehensive list
site models: discrete model (M3)

\[ P(x_h) = \sum_{i=0}^{K-1} p_i P(x_h \mid \omega_i) \]

Mixture-model likelihood calculation (see part 1)

\( \omega_0 = 0.01 \quad \omega_1 = 1.0 \quad \omega_2 = 2.0 \)
interpretation of a sites-model

\[ \omega_0 = 0.01 \quad \omega_1 = 1.0 \quad \omega_2 = 2.0 \]

diversifying selection (frequency dependent)
at 5% of sites with \( \omega_2 = 2 \)
models for variation among branches & sites

branch models
\( (\omega \text{ varies among branches}) \)

site models
\( (\omega \text{ varies among sites}) \)

branch-site models
\( (\text{combines the features of above models}) \)
models for variation among branches & sites

<table>
<thead>
<tr>
<th>variation ($\omega$) among branches &amp; sites:</th>
<th>approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang and Nielsen, 2002</td>
<td>fixed+mixture (ML)</td>
</tr>
<tr>
<td>Forsberg and Christiansen, 2003</td>
<td>fixed+mixture (ML)</td>
</tr>
<tr>
<td>Bielawski and Yang, 2004</td>
<td>fixed+mixture (ML)</td>
</tr>
<tr>
<td>Giundon et al., 2004</td>
<td>switching (ML)</td>
</tr>
<tr>
<td>Zhang et al. 2005</td>
<td>fixed+mixture (ML)</td>
</tr>
<tr>
<td>Kosakovskiy Pond et al. 2011, 2012</td>
<td>full mixture (ML)</td>
</tr>
</tbody>
</table>

*these methods can be useful when selection **pressures change over time at just a fraction of sites**

*it can be a challenge to apply these methods properly (**more about this later**)
branch-site “Model B”

MIXTURE-MODEL LIKELIHOOD

\[ P(x_h) = \sum_{i=0}^{K-1} p_i P(x_h | \omega_i) \]

\( \omega \) for background branches are from site-classes 1 and 2 (0.01 or 0.90)
two scenarios can yield branch-sites with $dN/dS > 1$

- 10% of sites have shifting balance on a fixed peak (same function)

- episodic adaptive evolution at 10% of sites for novel function

- branch-site codon models cannot tell which scenario is correct without external information!

Jones et al (2016) MBE
Jones et al (2018) MBE
"OMEGA MODELS"

\[ Q_{ij} = \begin{cases} 
0 & \text{if } i \text{ and } j \text{ differ by } > 1 \\
\pi_j & \text{for synonymous tv.} \\
\kappa \pi_j & \text{for synonymous ts.} \\
\omega \pi_j & \text{for non-synonymous tv.} \\
\omega k \pi_j & \text{for non-synonymous ts.} 
\end{cases} \]

Goldman and Yang (1994)
Muse and Gaut (1994)
3 analytical tasks

**task 1.** parameter estimation (e.g., $\omega$)

**task 2.** hypothesis testing

**task 3.** make predictions (e.g., sites having $\omega > 1$)
task 1: parameter estimation

Parameters: $t$ and $\omega$

Gene: acetylcholine $\alpha$ receptor

$\ln L = -2399$
task 1. parameter estimation (e.g., $\omega$)

task 2. hypothesis testing

LRT

task 3. prediction / site identification
task 2: likelihood ratio test for positive selection

\( H_0 \): variable selective pressure but NO positive selection (M1)
\( H_1 \): variable selective pressure with positive selection (M2)

Compare \( 2\Delta l = 2(l_1 - l_0) \) with a \( \chi^2 \) distribution

Model 1a (M1a)

\[ \hat{\omega} = 0.5 \quad (\omega = 1) \]

Model 2a (M2a)

\[ \hat{\omega} = 0.5 \quad (\omega = 1) \quad \hat{\omega} = 3.25 \]
task 3: identify the selected sites

task 1. parameter estimation (e.g., $\omega$) ✔

task 2. hypothesis testing ✔

task 3. prediction / site identification Bayes’ rule
task 3: which sites have $dN/dS > 1$

**model:**
9% have $\omega > 1$

**Bayes’ rule:**
site 4, 12 & 13

**structure:**
sites are in contact
review the mixture likelihood (model $M_3$)

$$P(x_h) = \sum_{i=0}^{K-1} p(\omega_i)P(x_h | \omega_i)$$

Total probability  Prior  Likelihood

$\omega_0 = 0.03$  $\omega_1 = 0.40$  $\omega_2 = 14.1$

$p_0 = 0.85$  $p_1 = 0.10$  $p_2 = 0.05$
Bayes' rule for identifying selected sites

Prior probability of hypothesis \((\omega_2)\)

\[
P(\omega_2 \mid x_h) = \frac{P(\omega_2) P(x_h \mid \omega_2)}{\sum_{i=0}^{K-1} P(\omega_i) P(x_h \mid \omega_i)}
\]

Likelihood of hypothesis \((\omega_2)\)

Site class 0: \(\omega_0 = .03\), 85% of codon sites

Site class 1: \(\omega_1 = .40\), 10% of codon sites

Site class 2: \(\omega_2 = 14\), 05% of codon sites

Posterior probability of hypothesis \((\omega_2)\)

Marginal probability (Total probability) of the data
task 3: Bayes rule for which sites have dN/dS > 1

NOTE: The posterior probability should NOT be interpreted as a “P-value”; it can be interpreted as a measure of relative support, although there is rarely any attempt at “calibration”.

- Site class 0: $\omega_0 = .03$ (strong purifying selection)
- Site class 1: $\omega_1 = .40$ (weak purifying selection)
- Site class 2: $\omega_2 = 14$ (positive selection)
Task 3: Bayes rule for which sites have $dN/dS > 1$

- **Naive Empirical Bayes**
  - Nielsen and Yang, 1998
  - Assumes no MLE errors

- **Bayes Empirical Bayes**
  - Yang et al., 2005
  - Accommodate MLE errors for some model parameters via uniform priors

- **Smoothed bootstrap aggregation**
  - Mingrone et al., MBE, 33:2976-2989
  - Accommodate MLE errors via bootstrapping
  - Ameliorates biases and MLE instabilities with kernel smoothing and aggregation
critical question:

Have the requirements for maximum likelihood inference been met?

(rarely addressed in real data analyses)
reguality conditions have been met

Normal MLE uncertainty (M2a)

- large sample size with regularity conditions
- MLEs approximately unbiased and minimum variance

\[ \hat{\theta} \sim N\left(\theta, I(\hat{\theta})^{-1}\right) \]
MLE instabilities (M2a)

- small sample sizes and $\theta$ on boundary
- continuous $\theta$ has been discretized (e.g., M2a)
- non-Gaussian, over-dispersed, divergence among datasets

bootstrapping can be used to diagnose this problem:


Mingrone et al., MBE, 33:2976-2989