part 3: analysis of natural selection pressure
phenomenological codon models do have many benefits:

- principled framework for statistical inference
- avoiding *ad hoc* corrections of “counting” methods
- computation of transition probabilities *
- explicit use of phylogeny
- model $\omega$ variation among sites
- model $\omega$ variation among branches
- many other kinds of models for $\omega$

* Computation of transition probabilities accomplishes, in just one step, (1) a proper correction for multiple substitutions, (2) weighting for alternative pathways between codons and (3) is the basis for estimating the values of the model parameters from the data in hand.
two basic types of models

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

site models
(\(\omega\) varies among sites)
branch models

<table>
<thead>
<tr>
<th>variation ((\omega)) 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
interpretation of a branch model

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

- useful when at some sites evolve under **diversifying selection** pressure over long periods of time
- this is not a comprehensive list

<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>
site models: discrete model (M3)

$$P(x_h) = \sum_{i=0}^{K-1} p_i P(x_h | \omega_i)$$
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
\( \omega \) varies among branches

site models
\( \omega \) varies among sites

branch-site models
(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>Kosakovsky 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

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

\( \omega = 0.01 \) \( \omega = 0.90 \) \( \omega = 5.55 \)

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

- $\omega = 0.01$ and $\omega = 0.90$ lead to different outcomes.
- $\omega_{FG} = 5.55$ indicates episodic adaptive evolution at 10% of sites for a novel function.
- 10% of sites have shifting balance on a fixed peak (same function).

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

- Jones et al. (2016) submitted: MBE.
model-based inference
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

t, κ, ω = unknown constants estimated by ML

π’s = empirical [GY: F3×4 or F61 in Lab]

use a numerical hill-climbing algorithm to maximize the likelihood function
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

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

H₀: variable selective pressure but NO positive selection (M1)
H₁: variable selective pressure with positive selection (M2)

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

Model 1a

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

Model 2a

\[ \hat{\omega} = 0.5 \quad (\omega = 1) \quad \hat{\omega} = 3.25 \]
task 2: likelihood ratio test for positive selection

**H₀**: Beta distributed variable selective pressure (M7)
**H₁**: Beta plus positive selection (M8)

Compare $2\Delta l = 2(l_1 - l_0)$ with a $\chi^2$ distribution.
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
Suppose that a population consists of 60% males and 40% females, and a disease occurs at the rate 1% in males and 0.1% in females.

\[ P(D) = P(M)P(D|M) + P(F)P(D|F) \]

Q₁: What is the probability that any individual carries the disease?

A₁: \[ 0.6 \times 0.01 + 0.4 \times 0.001 = 0.0064 \]

See Yang and Bielawski (2000) TREE 15:496-503 for a detailed presentation of this example.
$Q_2$: Given that an individual carries the disease, what is the probability that it is a male?

$A_2$: $0.6 \times 0.01 / 0.0064 = 0.94$

\[
P(M|D) = \frac{P(M) P(D|M)}{P(D)}
\]

See Yang and Bielawski (2000) TREE 15:496-503 for a detailed presentation of this example
Bayes’ rule in statistics

\[
Pr(\theta|D) = \frac{Pr(D|\theta) Pr(\theta)}{\sum_\theta Pr(D|\theta) Pr(\theta)}
\]

- **Likelihood** of hypothesis \(\theta\)
- **Prior probability** of hypothesis \(\theta\)
- **Posterior probability** of hypothesis \(\theta\)
- **Marginal probability** of the data (marginalizing over hypotheses)

from Paul Lewis’ lecture ....
1. analytical task-3

identifying selected sites under a codon model

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

Total probability  Prior  Likelihood

\[ \omega_0 = 0.03 \quad \omega_1 = 0.40 \quad \omega_2 = 14.1 \]

\[ p_0 = 0.85 \quad p_1 = 0.10 \quad p_2 = 0.05 \]
Bayes’ rule for identifying selected sites

Prior probability of hypothesis ($\omega_2$)

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

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 $dN/dS > 1$

![Graph showing posterior probabilities for different site classes with annotations]

- **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)

**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”.
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, under review
• accommodate MLE errors via bootstrapping
• ameliorates biases and MLE instabilities with kernel smoothing and aggregation

task 3: Bayes rule for which sites have dN/dS > 1
model based inference

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

task 2. hypothesis testing

task 3. prediction / site identification

let’s put this into practice …
example analysis (& experimental validation)
colour diversity of coral pigments (GFPs)

- Is color diversity tuned by natural selection?
- Is there a relationship between colour and endosymbiotic algae?

signal 1: long term (diversifying) selection

Bayes’ rule:

\[ P(\omega_2 | x_h) = \frac{p_2 P(x_h | \omega_2)}{P(x_h)} = \frac{p_2 P(x_h | \omega_2)}{\sum_{i=0}^{K-1} p_i P(x_h | \omega_i)} \]

signal 2: episodic selection

Bacteria were engineered to express the extant and ancestral GFP-like proteins. These bacteria were then cultured in a pattern that corresponded to the GFP-LIKE gene tree.

false biological conclusions
false biological conclusions

1. codon usage

2. process variation among sites

3. process variation over time

4. recombination

5. regularity conditions not met
how to model codon frequencies?

<table>
<thead>
<tr>
<th>Codon</th>
<th>Phe</th>
<th>Ser</th>
<th>Tyr</th>
<th>Cys</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TTC</td>
<td>27</td>
<td>15</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>TTA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TTG</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Codon</th>
<th>Leu</th>
<th>Pro</th>
<th>His</th>
<th>Arg</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAT</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TCA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TAG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TCA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CTC</td>
<td>2</td>
<td>15</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>CTA</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CTG</td>
<td>29</td>
<td>11</td>
<td>14</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Codon</th>
<th>Ile</th>
<th>Thr</th>
<th>Asn</th>
<th>Ser</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATT</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>ATC</td>
<td>12</td>
<td>11</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>ATA</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Met</td>
<td>4</td>
<td>4</td>
<td>37</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Codon</th>
<th>Val</th>
<th>Ala</th>
<th>Asp</th>
<th>Gly</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTT</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>GTC</td>
<td>2</td>
<td>38</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>GTA</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>GTG</td>
<td>25</td>
<td>3</td>
<td>30</td>
<td>0</td>
</tr>
</tbody>
</table>
how to model codon frequencies?

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>substitution rates are proportional to empirical frequency of:</td>
<td>target codon</td>
</tr>
<tr>
<td>target nucleotide</td>
<td></td>
</tr>
</tbody>
</table>

See Rodrique et al. (2008) for a comparison of GY and MG style codon models that suggests the MG style, combined with parameters for codon preferences, might be the most desirable core-model for future development.

The MutSel process (part 1) is inherently a process whereby the transition probability depends on the target nucleotide (MG).
how to model codon frequencies?

depending on the gene/gene, the method could yield **biased estimates of dN/dS**, See the following for cases:


example: A → C

AAA → CAA
AAA → ACA
AAA → AAC

<table>
<thead>
<tr>
<th>Δ at codon position</th>
<th>1\textsuperscript{st}</th>
<th>2\textsuperscript{nd}</th>
<th>3\textsuperscript{rd}</th>
</tr>
</thead>
<tbody>
<tr>
<td>GY</td>
<td>(\pi_{CA} )</td>
<td>(\pi_{CA} )</td>
<td>(\pi_{CA} )</td>
</tr>
<tr>
<td>MG</td>
<td>(\pi_{c1} )</td>
<td>(\pi_{c2} )</td>
<td>(\pi_{c3} )</td>
</tr>
</tbody>
</table>
false biological conclusions

1. codon usage

2. process variation among sites

3. process variation over time

4. recombination

5. regularity conditions not met
sequence evolution is complex

loop structures extend into extra-cellular space: **Hydrophilic amino acids here**

loop structures extend into cytoplasm: **Hydrophilic amino acids here**

loop structures span the membrane: **Hydrophobic amino acids here**

\[
\begin{align*}
\omega_0 & \quad \pi_0 \quad \kappa_0 \quad c_0 \\
\omega_1 & \quad \pi_1 \quad \kappa_1 \quad c_1 \\
\omega_2 & \quad \pi_2 \quad \kappa_2 \quad c_2
\end{align*}
\]

codon models: biological interpretation of differences among sites in \( \omega \) requires that such differences are due to selection pressure alone

GY-type codon models: variable \( \omega \)'s + \( c \)'s among sites = variable \( d_N \) & \( d_S \) among sites
modeling process variation among sites

<table>
<thead>
<tr>
<th>process variation among sites</th>
<th>software &amp; references</th>
</tr>
</thead>
<tbody>
<tr>
<td>• synonymous rate</td>
<td>several methods in:</td>
</tr>
<tr>
<td>• nonsynonymous rate</td>
<td><strong>HyPhy</strong>: Kosakovsky Pond et al. (2005)</td>
</tr>
<tr>
<td></td>
<td><strong>Datamonkey</strong>: Delport et al. (2010)</td>
</tr>
<tr>
<td>• baseline DNA/RNA substitution rate</td>
<td><strong>MultiLayer</strong>: Rubinstein et al. (2011)</td>
</tr>
<tr>
<td>• nonsynonymous rate</td>
<td><strong>LiBaC</strong>: Bao et al. (2008)</td>
</tr>
<tr>
<td>• baseline DNA/RNA substitution rate</td>
<td></td>
</tr>
<tr>
<td>• transition/transversion ratio</td>
<td></td>
</tr>
<tr>
<td>• codon frequencies</td>
<td></td>
</tr>
<tr>
<td>• nonsynonymous rate</td>
<td></td>
</tr>
</tbody>
</table>

several studies show **false signal for** $dN/dS > 1$ **is possible when process variation among sites in inadequately modeled**
false biological conclusions

1. codon usage

2. process varies among sites

3. process varies over time

4. recombination

5. regularity conditions not met
false signal for $dN/dS > 1$ is possible when codon frequencies change over time (non-stationarity). See Bay and Bielawski (2013) JME 76:205-15.
false biological conclusions

1. codon usage
2. variation among sites
3. variation over time
4. recombination
5. regularity conditions not met
High levels of recombination can yield false signal for \( dN/dS > 1 \) via the LRT. See...


For a nice solution see:

Note: Recombination adds among site variation relative to both process and phylogeny! See Sullivan et al. 2006 PLoS Biology 4: e234 for details.
false biological conclusions

1. codon usage

2. variation among sites

3. variation over time

4. recombination

5. regularity conditions not met
Normal MLE uncertainty (M2a)

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

\[ \hat{\theta} \sim N\left( \theta, I(\hat{\theta})^{-1} \right) \]
regularity conditions have **NOT** been met

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, under review
best practices
best practices in evolutionary surveys

1. processing and Q.C. (in large scale surveys)
2. alignment (independent evaluations)
3. recombination
4. robustness: MG vs GY style codon model
5. robustness: alternative tree topologies
6. robustness: variation in baseline DNA/RNA rates
7. bootstrapping

for discussion of best practices in large scale gene surveys see:

• Baker et al. (2016) Genetics, 203:905-22
nuclear receptor NR1D1: positive selection along human lineage?

1. alignment (independent evaluations) ✔
2. recombination ✔
3. robustness: MG vs GY ✔
4. robustness: tree topologies ✔
5. robustness: baseline DNA/RNA rates ✔
6. bootstrapping

instabilities in the MLEs

KEY

$\omega_M$: $\omega_{Mammal}$

$\omega_{GA}$: $\omega_{GreatApe}$

$\omega_{HC}$: $\omega_{Human-Chimpanzee}$

$\omega_H$: $\omega_{Human}$
What are the next steps in codon models?
What are the next steps in codon models?

1. applications of the MutSel framework
   - Tamuri AU et al. (2014) Genetics 197:257
   - Tamuri et al. (2012) Genetics 190:1101

2. joint modeling of genotype & phenotype
   - Nabholz et al. (2013) Genome Biol Evol 5:1273
THE END.