part 3: analysis of natural selection pressure

Q_{ij} = 0 if i and j differ by > 1 
\pi_i \text{ for synonymous \textsc{ts},} 
\kappa_i \text{ for synonymous \textsc{ts},} 
\sigma_i \text{ for non-synonymous \textsc{ts},} 
\kappa_{\text{ext}} \text{ for non-synonymous \textsc{ts},} 

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

“OMEGA MODELS”

\[
Q_i = \begin{cases} 
0 & \text{if } i \text{ and } j \text{ differ by } > 1 \\
\pi_j & \text{for synonymous tv.} \\
\kappa_j & \text{for synonymous ts.} \\
\omega_j & \text{for non-synonymous tv.} \\
\omega_b & \text{for non-synonymous ts.} 
\end{cases}
\]

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

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>Kosakovskyy 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
• useful when at some sites evolve under diversifying selection pressure over long periods of time

<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>

* this is not a comprehensive list

site models: discrete model (M3)

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

conditional likelihood calculation (see part 1)

ω₀ = 0.01  \hspace{0.5cm} ω₁ = 1.0  \hspace{0.5cm} ω₂ = 2.0
interpretation of a sites-model

5% of sites

ω₀ = 0.01  ω₁ = 1.0  ω₂ = 2.0

diversifying selection
(frequency dependent)
at 5% of sites with ω₂ = 2

models for variation among branches & sites

branch models
(ω varies among branches)

site models
(ω varies among sites)

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

<table>
<thead>
<tr>
<th>variation (ω) 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**

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

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

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

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) MBE
Jones et al. (2018) MBE

"Omega models"

\[
Q_{i,j} = \begin{cases} 
0 & \text{if } i \text{ and } j \text{ differ by } > 1 \\
\pi_i & \text{for synonymous } t_v \\
\kappa_{ij} & \text{for synonymous } t_s \\
\sigma_{ij} & \text{for non-synonymous } t_v \\
\omega_{ij} & \text{for non-synonymous } t_s \\
\end{cases}
\]

Goldman and Yang (1994)
Muse and Gaut (1994)
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 $\omega > 1$)

Parameter: $t$ and $\omega$  
Gene: acetylcholine $\alpha$ receptor  

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

task 2. hypothesis testing

$\text{LRT}$

task 3. prediction / site identification

---

**task 2: statistical significance**

**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  

---

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

**Bayes’ rule:**  
site 4, 12 & 13

**structure:**
sites are in contact

---

**task 3:** which sites have $dN/dS > 1$
review the mixture likelihood (model M3)

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

Prior probability of hypothesis \( (\omega) \)

Likelihood of hypothesis \( (\omega) \)

Posterior probability of hypothesis \( (\omega) \)

Marginal probability (Total probability) of the data

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

Site class 0: \( \omega_0 = .03 \), 85% of codon sites
Site class 1: \( \omega_1 = .40 \), 10% of codon sites
Site class 2: \( \omega_2 = 14.1 \), 05% of codon sites

\[ \omega_0 = 0.03 \quad \omega_1 = 0.40 \quad \omega_2 = 14.1 \]
\[ p_0 = 0.85 \quad p_1 = 0.10 \quad p_2 = 0.05 \]
task 3: Bayes rule for which sites have $dN/dS > 1$

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”.

Bayes’ rule

empirical Bayes

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

bootstrap
critical question:

Have the requirements for maximum likelihood inference been met?

(rarely addressed in real data analyses)

regularity conditions have been met

Normal MLE uncertainty (M2a)

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

\[ \hat{\theta} - N\left( \theta, \frac{1}{\hat{\sigma}^2} \right) \]
MLE instabilities (M2a)
- small sample sizes and $\beta$ on boundary
- continuous $\beta$ 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

part 4: phenomenological load and biological inference
review types of models

phenomenological

mechanistic

Newton
\[ F = \frac{G m_1 m_2}{r^2} \]

Einstein
\[ G_{\alpha\beta} = 8\pi T_{\alpha\beta} \]

molecular evolution is **process** and **pattern**

**process**  \[ \rightarrow \]  **pattern**

"MutSel models"

\[
Pr = \begin{cases} 
\mu_i N \times \frac{1}{N} = \mu_i & \text{if neutral} \\
\mu_i N \times \frac{2s_{ij}}{1 - e^{-2s_{ij}}} & \text{if selected} 
\end{cases}
\]

\[ s_{ij} = \Delta f_{ij} \]

Halpern and Bruno (1998)
Maximum phenomenological model for sequence data: explains all variation in a particular dataset

- so-called “saturated model” (multinomial model)
- does not generalize to other datasets
- no information about process
- highest lnL score (useless?)

**Question:** Does anyone really care, at all, that site pattern No.4 occurs 33 times in my sample of 5 mammalian mt genomes?

---

phenomenological load

**Maximum phenomenological model for sequence data**

- explains all variation in a particular dataset
  - so-called “saturated model” (multinomial model)
  - does not generalize to other datasets
  - no information about process
  - highest lnL score (useless?)

**Question:** Does anyone really care, at all, that site pattern No.4 occurs 33 times in my sample of 5 mammalian mt genomes?

---

**phenomenological load**

A different look at the issue ...

Kullback-Leibler divergence

\[
KL = \sum_x P_x(X | \hat{\theta}_x) \log \frac{P_T(X | \hat{\theta}_T)}{P_M(X | \hat{\theta}_M)}
\]

\[
P_T = \left( X | \hat{\theta}_T \right)
\]

true model (\(M_T\))

\[
P_M = \left( X | \hat{\theta}_M \right)
\]

fitted model (\(M_{\text{Poisson}}\))
**Not to scale!**

**Poisson model (M_p):**
- Single rate parameter

**Line:**
- Subspace

**Saturated model (M_s):**
- As many parameters as unique site patterns

"Deviance M_p:"

\[ D_{M_p} = -2 \left\{ \ell_{M_p}(\hat{\theta}_{M_p}|X,T) - \ell_{M_s}(X) \right\} \]

**LRR**

\[ \text{LRR} = D_{M0} - D_{M1} = -2 \left\{ \ell_{M1}(\hat{\theta}_{M1}|X,T) - \ell_{M2}(\hat{\theta}_{M2}|X,T) \right\} \]

**KEY POINT:** That addition of any parameter will reduce the deviance
The likelihood ratio test (LRT) is used to evaluate the significance of the additional parameters in M1.

\[ \text{LRR} = D_{M0} - D_{M1} = -2 \left\{ \ell_{M1}(\hat{\theta}_{M1}|X,T) - \ell_{M2}(\hat{\theta}_{M2}|X,T) \right\} \]

The LRT assumes that the null model is true \((M0 = M_T)\).

The null model is not on the path to the true model.

The problem:

IF…

• the null model fails to explain some fraction of site pattern variation generated by \(M_T\)

• the additional parameter(s) represent a false process but are confounded with \(M_T\)

THEN…

• the added parameters can be statistically significant even when they represent a false process

• the degree to which a significant parameter carries false information about process is called Phenomenological Load (PL)
Hypothesis tests along THIS PATH can have phenomenological load

- significant LRTs b/c variation is not random
- interpretation is not direct about mechanism of evolution

---

**RESEARCH ARTICLE**

Superiority of a mechanistic codon substitution model even for protein sequences in Phylogenetic analysis

Satoshi Miyakawa

---

**On the Need for Mechanistic Models in Computational Genomics and Metagenomics**

David A. Liberles, Ashley I. Teufel, Liang Lu, and Tanja Stadler

Department of Molecular Biology, University of Wyoming

Department of Statistics and Institute of Bioinformatics, University of Georgia

Institut für Integrative Biologie, Eidgenössische Technische Hochschule Zürich, Zürich, Switzerland

---

**A Generalized Mechanistic Codon Model**

Maryam Zaheri, Linda Dib, and Nicolas Salamin

Department of Ecology and Evolution, Biophore, University of Lausanne, 1015 Lausanne, Switzerland

Swiss Institute of Bioinformatics, Genopole, Quartier Sorge, 1015 Lausanne, Switzerland

These authors contributed equally to this work.
“Phase 1 thinking” develop and test models as if on this path

- assume null model as true
- ignore site pattern distributions
- believe in pure tests of mechanism
- treat model selection as if it’s modeling the data

“Phase 2 thinking” develop and test models as if on this path

- realistic model miss-specification
- assess confounding
- assess phenomenological load

Let’s take a look at an example of “Phase 2 thinking”
Percent Reduction in Deviance (PDR)

\[ \text{PRD} = \frac{D_{M0} - D_{M1}}{D_{\text{Poisson}}} \]

DT: Double and Triple mutations

Example double: ATG (Met) → AAA (Lys) [\( \alpha \) parameter]
Example triple: AAA (Lys) → GGG (GLY) [\( \beta \) parameter]

**M0 Q matrix**
- 2 parameters (\( \kappa \) and \( \omega \))
- DT not allowed

**New Q matrix**
- 4 parameters (\( \kappa \), \( \omega \), \( \alpha \), \( \beta \))
- DT allowed (via \( \alpha \) and \( \beta \))

Is such a model warranted?
Let's do a simulation study!

**process** ($M_f$):
- chimps
- gorilla
- orangutan
- Sumatran orangutan
- common gibbon
- harbor seal
- grey seal
- cat
- horse
- Indian rhinoceros
- cow
- fin whale
- blue whale
- rat
- mouse
- wallaroo
- opossum
- platypus

**outcome** ($X$):
- we need outcomes to match up

**simulation**
- MutSel
- $f^\theta$ differ for each site
- NO DT-mutations
- 12 mt proteins (3331 codons)
- 20 mammals

Our simulated data LOOKS LIKE the REAL DATA!

**simulation for** $M_f$:
- MutSel with NO DT-mutations

since there are NO DT-mutations, PRD is a measure of PL
Conclusions:
- DT parameters (α and β) carry PL
- there is evidence for DT process in mtDNA in excess of PL
- estimated level of DT very small in the real data

phenomenological load

testing PL on three proposed mechanisms for mtDNA

proposed evolutionary mechanisms

<table>
<thead>
<tr>
<th></th>
<th>DT mutations</th>
<th>relaxed selection</th>
<th>synonymous rate variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRD(dS)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRD((\alpha, \beta))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRD((\hat{k}))</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

PRD for real mtDNA dataset
**Why should you care?**

1. All of molecular evolution depends on models to some extent.
2. All models are underspecified.
3. Model parameters can carry substantial PL.
4. Faster computers ➔ more complex models
5. Next Gen sequencing ➔ minor effects detectable
6. Standard model selection tools will NOT inform you about levels of PL.
7. Excessive PL will lead to false biological conclusions.
8. Modelers MUST have biological expertise, and they MUST use that expertise as part of the modeling process.

---

**Shifting Balance and Phenomenological Load: References**


How can you really tell if you have learned anything relevant to the function of your protein?

- formally combine computational and experimental approaches (B. Chang, next lecture)
- formally combine phenotypic information within the computational analysis of sequence evolution

THE END.