part 3: analysis of natural selection pressure

\[ \Omega = \begin{cases} 0 & \text{if } i \neq j \text{ and } |i - j| > 1 \\ \pi_i & \text{for synonymous } i \\ \kappa_i & \text{for synonymous } i \\ \omega_i & \text{for non-synonymous } i \end{cases} \]

"OMEGA MODELS"

Goldman and Yang (1994)
Muse and Gaut (1994)
This codon model "M0"

\[
\omega_i = \begin{cases} 
0 & \text{if } i \text{ and } j \text{ differ by } > 1 \\
\pi_f & \text{for synonymous tv.} \\
\kappa_f & \text{for synonymous ts.} \\
\kappa_i & \text{for non-synonymous tv.} \\
\omega_i \kappa_f & \text{for non-synonymous ts.} 
\end{cases}
\]


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 ($\omega$) 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
site models*  

<table>
<thead>
<tr>
<th>Variation ($\omega$) 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;</td>
<td>mixture (Bayesian)</td>
</tr>
<tr>
<td>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 some sites evolve under diversifying selection pressure over long periods of time  
* this is not a comprehensive list

site models: discrete model (M3)

Mixture-model likelihood

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

conditional likelihood calculation (see part 1)

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

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>covarion-like (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>
<tr>
<td>Jones et al., 2016, 2018</td>
<td>covarion-like (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”

\[
P(x_h) = \sum_{i=0}^{K-1} p_i P(x_h | \omega)
\]
two scenarios can yield branch-sites with dN/dS > 1

model-based inference

“OMEGA MODELS”

\[ \omega = \begin{cases} 
0 & \text{if } i \neq j \text{ and } d_i > d_j \\
\pi & \text{for synonymous site} \\
\kappa & \text{for non-synonymous site} \\
\lambda & \text{for site with synonymous and non-synonymous variation} \\
\nu & \text{for site with synonymous variation} \\
\end{cases} \]

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 $\omega > 1$)

---

**task 1: parameter estimation**

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

Genealogy:
- Mouse
- Human
- Common ancestor

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

task 2. hypothesis testing **LRT**

task 3. prediction / site identification

task 2: statistical significance

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**)

Model 2a (**M2a**)

$\hat{\omega} = 0.5$  ($\omega = 1$)

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

- task 1. parameter estimation (e.g., $\omega$) ✔
- task 2. hypothesis testing ✔
- task 3. prediction / site identification Bayes’ rule

**model:**
10% have $\omega > 1$

**Bayes’ rule:**
site 4, 12 & 13

**structure:**
sites are in contact

**task 3: which sites have $dN/dS > 1$**
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 ($\omega_2$)

Posterior probability of hypothesis ($\omega_2$)

Marginal probability (Total probability) of the data

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
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".
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

\( \hat{\theta} \sim \mathcal{N}(\theta, I(\hat{\theta})^{-1}) \)
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

regularity conditions have NOT been met

software for codon models in the ML framework

PAML: a package of programs for process modeling

HyPhy: comparative sequence analysis using stochastic evolutionary models; http://www.hyphy.org/

DataMonkey: a server that supports a variety of HYPHY tools at no cost; http://www.datamonkey.org/

COLD: a program that implements a general-purpose parametric (GPP) codon model. Most codon models are special cases of the GPP codon model. https://github.com/ijk23/COLD


ModL: a program for restoring regularity when testing for positive selection using codon models https://github.com/Jehops/codeml_modl
part 4: phenomenological load and biological inference

Review types of models

Phenomenological load

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"**

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

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

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?)

**site pattern**

<table>
<thead>
<tr>
<th>site pattern</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTG CGT TCT</td>
<td>CCA GCC GAC AAC ACC GTC AAG GCC GGG TGG GGC AAG GCT GGC GCG CAC</td>
</tr>
</tbody>
</table>
| ... ... ... ... | ... ... ... ... ... ... ... ... ... ... | ...
| G.C ... T ... T ... T ... A ... A.T ... AA ... A.C ... A.GC ... ... |
| ... ... ... G ... A ... AT ... A ... A.A ... A.A ... A.GT ... G ... A ... T ... GC ... T ... |
| ... ... ... C ... G ... A ... AT ... A ... A.A ... A.A ... A.GT ... G ... A ... T ... GC ... T ... |

**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 …

\[ P_T = \{ X \theta_T \} \]
true model (Mᵀ)

\[ P_M = \{ X \theta_M \} \]

fitted model (Poisson)

\[ KL = \sum_{x} P_T(x \theta_T) \log \frac{P_T(x \theta_T)}{P_M(x \theta_M)} \]
Kullback-Leibler divergence

Not to scale!

Poisson model (M_p):
single rate parameter

Line:
subspace

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

\[ D_{M_p} = -2 \left\{ \ell_M(\theta_M, X, T) - \ell_M(X) \right\} \]
“Deviance M_p”
M1 extends M0 by the addition of parameters

**KEY POINT:** addition of any parameter will reduce the deviance

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

The **Likelihood Ratio Test (LRT)** "manages" phenomenological variability (not mechanistic variability)
let's do a simulation study

and

let's use “double mutations” and “triple mutations” as an example

element double (D): \text{ATG (Met) } \rightarrow \text{AAA (Lys)}

element triple (T): \text{AAA (Lys) } \rightarrow \text{GGG (GLY)}

the simulation and the outcomes...

process \( M_h \):
- MutSel
- \( f \) differ for each site
- NO DT-mutations
- 12 mt proteins (3331 codons)
- 20 mammals

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

Our simulated data LOOKS LIKE the REAL DATA!
DT: Double and Triple mutations

Example double: ATG (Met) → AAA (Lys) [α parameter]
Example triple: AAA (Lys) → GGG (GLY) [β parameter]

M0 Q matrix
- 2 parameters (κ and ω)
- DT not allowed

New Q matrix: M0 + DT
- 4 parameters (κ, ω, α, β)
- DT allowed (via α and β)

How to test such a model?

Percent Reduction in Deviance (PDR)

\[ \text{PRD} = \frac{D_{H_{null}} - D_{H_{alt}}}{D_{Poisson}} \]
simulation for $M_1$: MutSel with NO DT-mutations

since there are NO DT-mutations, PRD gives us a distribution for Phenomenological Load

phenomenological load
testing PL on three proposed mechanisms for mtDNA

proposed evolutionary mechanisms

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

PRD for real mtDNA dataset
part 5: re-assessing long-held paradigms for evidence of adaptive evolution

Three paradigms for “this” side:

1. codon substitution model: $d_N/d_S > 1$ is evidence of adaptive evolution

2. “mechanistic” substitution models are better

3. It’s easy to test and predict model performance via simulation
**Paradigm 1:** \( \frac{d_f}{d_s} > 1 \) is evidence of adaptive evolution of function
the MutSel fitness landscape: **adaptive evolution**

key result 1:
adaptive evolution: $p_+ > p_-$
("peak shift")

$d_n/d_S > 1$ (transient)

---

the MutSel fitness landscape: **non-adaptive shifting balance**

key result 2:
purifying selection: $p_+ = p_-$
(static landscape)

$d_n/d_S > 1$ (transient)
Paradigm 1: adaptive evolution

Realty: $d_N/d_S > 1$ on a fixed landscape with no change in function

Proposal: develop new frameworks that do NOT depend on the $d_N/d_S > 1$ paradigm

Paradigm 2: "mechanistic" substitution models should be better
All imply you move closer to a true mechanistic model

The Likelihood Ratio Test (LRT) "manages" phenomenological variability (not mechanistic variability)
Paradigm 2: *Myth Busted*

For real data: mechanistic parameters within models are expected to carry some **Phenomenological Load**

Proposal: intentionally add phenomenological parameters that improve inferences (e.g., covarion $\delta$)

Paradigm 3: It’s easy to test and predict model performance via simulation
Definable models

Simulating models

Inference models

Phenomenological codon models

Natural processes

Testing codon models

Definable models

Simulating models

Inference models

Natural processes

Slide from Michael Landis (adapted)
Paradigm 3: It’s easy to test model performance.

**In reality it’s hard to** (1) compare complex site pattern distributions, and (2) identify models that produce biologically plausible distributions

**Proposal:** we need to do more work on how to generate “realistic” site pattern distributions and change the way we think about testing model performance.
How can you really tell if you have learned anything relevant to the function of your protein?

• combine computational and experimental approaches (B. Chang, next lecture; “Gold Standard”)

• informal cross-validation via comparison with external phenotypic information (B. Chang, next lecture)

• formally include phenotypic information within the likelihood inference framework (we have this working; the paper is in revision... “stay tuned”)

THE END.