part 4: phenomenological load and biological inference

review types of models

phenomenological load

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

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

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

**process**

<table>
<thead>
<tr>
<th>“MUTSel models”</th>
</tr>
</thead>
</table>
| \[
Pr = \begin{cases} 
\mu_i \times \frac{1}{N} = \mu_i & \text{if neutral} \\
\mu_i N \times \frac{2n_i}{1 - e^{-2n_i}} & \text{if selected} 
\end{cases}
\] |

\[s_{ij} = \Delta_{ij}\]

Halpern and Bruno (1998)

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?)

**site pattern**

<table>
<thead>
<tr>
<th>site pattern</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCG GCC AAG AAC GAC GTG CAG GCC GTT GSG AAG GTT GCC GCG CAC</td>
<td></td>
</tr>
<tr>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>C . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>G . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td></td>
</tr>
</tbody>
</table>

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

“The good”
- all we have to model are “outcomes” (site pattern distribution)
- they can be predictive (e.g., Newtonian models)
- they can tell us about process (e.g., some codon models)

“The bad”
- a “saturated model” is useless
- must “decide” how much variability to “soak up” with model parameters
- matching variability to mechanistic process is hard
- traditional statistical methods manage phenomenological variability
  (NOT process variability)

“the ugly”
- getting it wrong = false biological conclusions

new concept: move phenomenological from model to parameter

phenomenological load (PL): if a parameter has a mechanistic interpretation, and if the process it represents did not actually occur, then when it absorbs significant variance that parameter has taken on phenomenological load (measured via PRD*).

two conditions for PL:
1. confounding of model parameters
2. underspecified model

* PRD = percent reduction of deviance, and is defined in subsequent slides
phenomenological load

codon models

1. confounding

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

DNA sub-model:
- \( \kappa \) and \( \pi \)
- Applied to all sites equally
- Not mutation sub-model

Protein level sub-model:
- \( \omega \) and \( \pi \)
- Direct selective interpretation
- Affected by mutation process

1. sub-models are confounded!

2. underspecified

missing model variability:
- Different fitness landscapes for sites
- Different AA exchangeabilities \( [s_{ij}] \)
- Different equilibrium for sites
- Independent mutational sub-model
- Mechanistic effect of \( N_e \)
- High level non-independence (global epistasis for stability)
- Low-level non-independence (local epistasis for function)

2. models are heavily underspecified

phenomenological load

a different look at the issue ...

true model (\( M_T \))

fitted model (\( M_0 \))
Kullback-Leibler divergence

\[ KL = \sum_{\tau} p_\tau (X | \theta_\tau) \log \frac{p_\tau (X | \theta_\tau)}{p_{M0} (X | \theta_{M0})} \]

"Deviance M0"

\[ D_{M0} = -2 \left\{ l_{M0} (\bar{\theta}_{M0}, X, T) - l_{M0} (X) \right\} \]
Percent Reduction in Deviance (PDR)

$$PRD = \frac{D_{M0} - D_{M3}}{D_{\text{poisson}}}$$

Hypothesis tests along THIS PATH have direct connection to mechanism of evolution

Hypothesis tests along THIS PATH have phenomenological load

- significant LRTs b/c variation is not random
- interpretation is not direct about mechanism of evolution
**DT: Double and Triple mutations**

- Example double: ATG (Met) $\Rightarrow$ AAA (Lys) \([\alpha \text{ parameter}]\)
- Example triple: AAA (Lys) $\Rightarrow$ GGG (GLY) \([\beta \text{ parameter}]\)

**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!

**Simulation** ($M_t$):
- MutSel
- $\beta^t$ 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!
simulation for $M_3$:
- MutSel with NO DT-mutations

since there are NO DT-mutations, PRD is a measure of PL

Conclusions:
- DT parameters ($\alpha$ and $\beta$) carry $\text{PL}$
- is evidence for DT process in mtDNA in excess of PL
- estimated level of DT very small in the real data
Why should you care?

1. All of molecular evolution depends on models to some extent.
2. All models are wrong (underspecified).
3. Model parameters will carry some 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.
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.