codon substitution models and the analysis of natural selection pressure

Joseph P. Bielawski
Department of Biology
Department of Mathematics & Statistics
Dalhousie University

introduction

codon models have many uses

- investigate process of protein evolution [this talk]
- phylogenetic inference
- ancestral sequence reconstruction
- dating divergence events
- alignment
- simulation
introduction

morphological adaptation

introduction

metabolic networks
protein structure

Troponin C: fast skeletal

Troponin C: cardiac and slow skeletal

introduction

gene sequences

human
cow
rabbit
rat
opossum

GTG CTG TCT CCT GCC GAC ACC AAC GTC AAG GCC GCC TGG GCC AAG GTT GCC GCG CAC

... ... .G.C ... ... T. ...T ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ......
“there is no single statistic which is best for testing the most general departures from neutrality” (Watterson 1977)

Powerful analytical tools:
1. Population genetic data
2. Comparative analysis of codon sequences
3. Comparative analysis of amino acid sequences

1. The $\omega$ ratio ($d_N/d_S$): the basics
2. Markov models of codon evolution
3. Model based inference
4. Sequence evolution is complex ("know your data")
5. PAML introduction
6. Real data exercises
The genetic code determines how random changes to the gene brought about by the process of mutation will impact the function of the encoded protein.

Codon models treat codons as the independent units, not individual nucleotide sites.

\[ d_S: \] number of synonymous substitutions per synonymous site \((K_S)\)

\[ d_N: \] number of nonsynonymous substitutions per nonsynonymous site \((K_A)\)

\[ \omega: \] the ratio \(d_N/d_S\); it measures selection at the protein level

Kimura (1968)

The basics

Why use \(d_N\) and \(d_S\)?
(Why not use raw counts?)

Example of counts:

- 300 codon gene from a pair of species
- 5 synonymous differences
- 5 nonsynonymous differences

\[ 5/5 = 1 \]

Why don’t we conclude that rates are equal (i.e., neutral evolution)?
Relative proportion of different types of mutations in hypothetical protein coding sequence.

<table>
<thead>
<tr>
<th>Type</th>
<th>Expected number of changes (proportion)</th>
<th>All 3 Positions</th>
<th>1st positions</th>
<th>2nd positions</th>
<th>3rd positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total mutations</td>
<td></td>
<td>549 (100)</td>
<td>183 (100)</td>
<td>183 (100)</td>
<td>183 (100)</td>
</tr>
<tr>
<td>Synonymous</td>
<td></td>
<td>134 (25)</td>
<td>8 (4)</td>
<td>0 (0)</td>
<td>126 (69)</td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td></td>
<td>392 (71)</td>
<td>166 (91)</td>
<td>176 (94)</td>
<td>57 (27)</td>
</tr>
<tr>
<td>nonsense</td>
<td></td>
<td>23 (4)</td>
<td>9 (5)</td>
<td>7 (4)</td>
<td>7 (4)</td>
</tr>
</tbody>
</table>

Modified from Li and Graur (1991). Note that we assume a hypothetical model where all codons are used equally and that all types of point mutations are equally likely.

---

Why use $d_N$ and $d_S$?

Same example, but using $d_N$ and $d_S$:

- Synonymous sites = 25.5%
  
  \[ S = 300 \times 3 \times 25.5\% = 229.5 \]

- Nonsynonymous sites = 74.5%
  
  \[ N = 300 \times 3 \times 74.5\% = 670.5 \]

So, $d_S = 5/229.5 = 0.0218$

\[ d_N = 5/670.5 = 0.0075 \]

\[ d_N/d_S (\omega) = 0.34, \text{ purifying selection} \text{ !!!} \]
Relative proportion of different types of mutations in hypothetical protein coding sequence.

<table>
<thead>
<tr>
<th>Type</th>
<th>All 3 Positions</th>
<th>1st positions</th>
<th>2nd positions</th>
<th>3rd positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total mutations</td>
<td>549 (100)</td>
<td>183 (100)</td>
<td>183 (100)</td>
<td>183 (100)</td>
</tr>
<tr>
<td>Synonymous</td>
<td>134 (25)</td>
<td>8 (4)</td>
<td>0 (0)</td>
<td>126 (69)</td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>392 (71)</td>
<td>166 (91)</td>
<td>176 (94)</td>
<td>57 (27)</td>
</tr>
<tr>
<td>nonsense</td>
<td>23 (4)</td>
<td>9 (5)</td>
<td>7 (4)</td>
<td>7 (4)</td>
</tr>
</tbody>
</table>

Modified from Li and Graur (1991). Note that we assume a hypothetical model where all codons are used equally and that all types of point mutations are equally likely.

Note: by framing the counting of sites in this way we are using a “mutational opportunity” definition of the sites. Not everyone agrees that this is the best approach. For an alternative view see Bierne and Eyre-Walker 2003 Genetics 168:1587-1597.

Partial codon usage table for the GstD1 gene of Drosophila

<table>
<thead>
<tr>
<th></th>
<th>Phe</th>
<th>Ser</th>
<th>Thr</th>
<th>Tyr</th>
<th>Trp</th>
<th>Cys</th>
<th>Glu</th>
<th>Gln</th>
<th>Lys</th>
<th>His</th>
<th>Arg</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TTC</td>
<td>27</td>
<td>15</td>
<td>22</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TTA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TCG</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TCG</td>
<td>15</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TAA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TGA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TGG</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TAT</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TAC</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TGT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CTA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CTC</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CGG</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CGA</td>
<td>15</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CCG</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CCC</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CCA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CGT</td>
<td>29</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CGC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CCG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

transitions vs. transversions:  
\( \frac{\text{ts}}{\text{tv}} = 2.71 \)

preferred vs. un-preferred codons:
1. Count synonymous (S) and nonsynonymous (N) sites
2. Count synonymous and nonsynonymous differences
3. Apply some corrections (e.g., correct for multiple hits)

---

**selected counting methods**

"corrections" and estimation bias in $d_s$

Data from: Dunn, Bielawski, and Yang (2001) Genetics, 157: 295-305

---

Model based inference:

1. assumptions are explicit
2. corrections are not ad hoc
3. explicit use of a phylogeny improves power
4. principled framework for modelling and inference of the biology

Modelling decisions are just as important (assumptions often matter more than methods!)
"A model is an intentional simplification of a complex situation designed to eliminate extraneous detail in order to focus attention on the essentials of the situation" (Daniel L. Hartl)
important parameters:

- transition/transversion rate ratio: $\kappa$
- biased codon usage: $\pi_j$ for codon $j$
- nonsynonymous/synonymous rate ratio: $\omega = d_N / d_S$

rates to CTG

**Synonymous**

- $\text{CTC (Leu)} \rightarrow \text{CTG (Leu)}$: $\pi_{\text{CTG}}$
- $\text{TTC (Leu)} \rightarrow \text{CTG (Leu)}$: $\kappa \pi_{\text{CTG}}$

**Nonsynonymous**

- $\text{GTG (Val)} \rightarrow \text{CTG (Leu)}$: $\omega \pi_{\text{CTG}}$
- $\text{CCG (Pro)} \rightarrow \text{CTG (Leu)}$: $\kappa \omega \pi_{\text{CTG}}$
"GY-style" codon models

<table>
<thead>
<tr>
<th>from codon below:</th>
<th>TTT (Phe)</th>
<th>TIC (Phe)</th>
<th>TTA (Leu)</th>
<th>TTG (Leu)</th>
<th>CIT (Leu)</th>
<th>CTC (Leu)</th>
<th>GGG (Gly)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT (Phe)</td>
<td>—</td>
<td>$\kappa_{TIC}$</td>
<td>$\omega \pi_{TIC}$</td>
<td>$\omega \pi_{TTG}$</td>
<td>$\omega \pi_{TTL}$</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td>TTC (Phe)</td>
<td>$\alpha_{TTT}$</td>
<td>—</td>
<td>$\omega \pi_{TIC}$</td>
<td>$\omega \pi_{TTG}$</td>
<td>0</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td>TTA (Leu)</td>
<td>$\omega \pi_{TIC}$</td>
<td>$\omega \pi_{TIC}$</td>
<td>—</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>TTG (Leu)</td>
<td>$\omega \pi_{TIC}$</td>
<td>$\omega \pi_{TIC}$</td>
<td>$\alpha_{TTA}$</td>
<td>—</td>
<td>0</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td>CIT (Leu)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$\omega \pi_{CTC}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CTC (Leu)</td>
<td>$\omega \pi_{CTT}$</td>
<td>0</td>
<td>0</td>
<td>$\omega \pi_{CTC}$</td>
<td>—</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GGG (Gly)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>—</td>
</tr>
</tbody>
</table>

* This is equivalent to the codon model of Goldman and Yang (1994). Parameter $\omega$ is the ratio $d_N/d_S$, $\kappa$ is the transition/transversion rate ratio, and $\pi_i$ is the equilibrium frequency of the target codon (i).

$$P(t) = \{p_{ij}(t)\} = e^{Qt}$$

---

"GY-style" codon models

(Goldman & Yang 1994 MBE 11:725-736)

$$q_{ij} = \begin{cases} 
0 & \text{if } i \text{ and } j \text{ differ at 2 or 3 positions} \\
\pi_j, & \text{for syn. transversion} \\
K\pi_j, & \text{for syn. transition} \\
\omega \pi_j, & \text{for nonsyn. transversion} \\
\omega K\pi_j, & \text{for nonsyn. transition}
\end{cases}$$

$$P(t) = \{p_{ij}(t)\} = e^{Qt}$$
The likelihood of observing the entire sequence alignment is the product of the probabilities at each site.

\[ L = L_1 \times L_2 \times L_3 \times \ldots \times L_N = \prod_{h=1}^{N} L_h \]

The log likelihood is a sum over all sites.

\[ \ell = \ln \{ L \} = \ln \{ L_1 \} + \ln \{ L_2 \} + \ln \{ L_3 \} + \ldots + \ln \{ L_N \} = \sum_{h=1}^{N} \ln \{ L_h \} \]
the good: we now have a framework for …
  o avoiding *ad hoc* corrections of counting methods
  o computation of transition probabilities *
  o principled framework for statistical inference

a new issue: averaging ω over a pair of sequences has very low power to detect positive selection if the question is about *“when”* or *“where”* ω > 1!

* 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.

Our question: *When?*
If we average over the tree, we do NOT detect positive selection; \( \omega = 0.49 \).

Grey branches: \( \omega = 0.2 \)
Black branches: \( \omega = 0.5 \)
Blue branches: \( \omega = 1.2 \)

Our question: Where?
If we average over sites, we do NOT detect positive selection: \( \omega = 0.31 \)

- Purifying: \( d_\text{N}/d_\text{S} = 0.01 \)
- Neutral: \( d_\text{N}/d_\text{S} = 1 \)
- Adaptive: \( d_\text{N}/d_\text{S} = 2 \)

**Problem:** averaging \( \omega \) over a pair has very low power if the questions are about "when" or "where"!

**Solution:** phylogenetic estimation of selection pressure
L(x_k) = \sum_k \prod p_{kx_k}(t_k) p_{kx_j}(t_j)

Sum over all possible codons of ancestral nodes (TTT, TTC, ..., GGG)

We made some more progress ....

1. Branch models
   \( (\omega \text{ varies among branches}) \)

2. Site models
   \( (\omega \text{ varies among sites}) \)

3. Branch-site models
   \( (\text{combines the features of above models}) \)
These methods can be useful when selection pressure is strongly episodic.

Variation (ω) among branches:

<table>
<thead>
<tr>
<th>Approach</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.*
1. branch models

- species colonization of a new niche
- altered context for gene expression
- gene duplication event(s)
- lateral gene transfers (LGTs)
- cross-species virus transmission & host switching
- organismal adaptive radiations

2. site models *

Variation (\( \omega \)) among sites:

<table>
<thead>
<tr>
<th>Approach</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>fixed effects (ML)</td>
<td>Yang and Swanson, 2002</td>
</tr>
<tr>
<td>fixed effects (ML)</td>
<td>Bao, Gu and Bielawski, 2006</td>
</tr>
<tr>
<td>site wise (LRT)</td>
<td>Massingham and Goldman, 2005</td>
</tr>
<tr>
<td>site wise (LRT)</td>
<td>Kosakovský Pond and Frost, 2005</td>
</tr>
<tr>
<td>mixture model (ML)</td>
<td>Nielsen and Yang, 1998</td>
</tr>
<tr>
<td>mixture model (ML)</td>
<td>Kosakovský Pond, Frost and Muse, 2005</td>
</tr>
<tr>
<td>mixture (Bayesian)</td>
<td>Huelsenbeck and Dyer, 2004</td>
</tr>
<tr>
<td>mixture (Bayesian)</td>
<td>Huelsenbeck et al. 2006</td>
</tr>
<tr>
<td>mixture (LiBaC/MBC)</td>
<td>Bao, Gu, Dunn and Bielawski 2008</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 (over a phylogeny):
1. fixed effect codon models
2. site-specific methods
3. random effect codon models ("M-series" models*)
4. LiBaC (Hard & Soft Model Based Clustering)

* we will review a very select set of examples

---

2. site models: “M-series”

<table>
<thead>
<tr>
<th>Model</th>
<th>Code</th>
<th>NP</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-ratio</td>
<td>M0</td>
<td>1</td>
<td>$\omega$</td>
</tr>
<tr>
<td>Neutral</td>
<td>M1a</td>
<td>2</td>
<td>$p_0$, $\alpha_0$</td>
</tr>
<tr>
<td>Selection</td>
<td>M2a</td>
<td>4</td>
<td>$p_0$, $\alpha_0$, $\alpha_2$</td>
</tr>
<tr>
<td>Discrete</td>
<td>M3</td>
<td>2K-1</td>
<td>$p_0$, $\alpha_0$, $\alpha_2$</td>
</tr>
<tr>
<td>Frequency</td>
<td>M4</td>
<td>5</td>
<td>$p_0$, $\alpha_0$, $\alpha_2$</td>
</tr>
<tr>
<td>Gamma</td>
<td>M5</td>
<td>2</td>
<td>$\alpha$, $\beta$</td>
</tr>
<tr>
<td>2Gamma</td>
<td>M6</td>
<td>4</td>
<td>$p_0$, $\alpha_0$, $\beta_1$, $\alpha_1$</td>
</tr>
<tr>
<td>Beta</td>
<td>M7</td>
<td>2</td>
<td>$p$, $q$</td>
</tr>
<tr>
<td>Beta&amp;$\omega$</td>
<td>M8</td>
<td>4</td>
<td>$p_0$, $p$, $q$, $\omega$</td>
</tr>
<tr>
<td>Beta&amp;gamma</td>
<td>M9</td>
<td>5</td>
<td>$p_0$, $p$, $q$, $\alpha$, $\beta$</td>
</tr>
<tr>
<td>Beta&amp;normal+1</td>
<td>M10</td>
<td>5</td>
<td>$p_0$, $p$, $q$, $\alpha$, $\beta$</td>
</tr>
<tr>
<td>Beta&amp;normal+1</td>
<td>M11</td>
<td>5</td>
<td>$p_0$, $p$, $q$, $\mu$, $\sigma$</td>
</tr>
<tr>
<td>0&amp;2normal+1</td>
<td>M12</td>
<td>5</td>
<td>$p_0$, $p_1$, $\mu_2$, $\alpha_1$, $\alpha_2$</td>
</tr>
<tr>
<td>3normal+0</td>
<td>M13</td>
<td>6</td>
<td>$p_0$, $p_1$, $\mu_2$, $\alpha_1$, $\alpha_2$, $\sigma_1$</td>
</tr>
</tbody>
</table>
2. site models: \( \omega \) varies among sites

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

\( \omega = 0.01 \quad \omega = 1.0 \quad \omega = 2.0 \)
o vaccine design
o genetic incompatibilities in human infertility
o non-hormonal contraception drugs
o identify pathogenicity genes
o identify candidate genes for drug therapies
o identify immune and defense system genes
o aid functional classification of unknown genes
o incorporate in models of protein 2D and 3D structure

2. site models

- codon models

<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 (ML)</td>
</tr>
<tr>
<td>Forsberg and Christiansen, 2003</td>
<td>fixed (ML)</td>
</tr>
<tr>
<td>Bielawski and Yang, 2004</td>
<td>fixed (ML)</td>
</tr>
<tr>
<td>Giundon et al., 2004</td>
<td>switching (Bayesian)</td>
</tr>
<tr>
<td>Zhang et al. 2005</td>
<td>mixture (ML)</td>
</tr>
<tr>
<td>Kosakovskv Pond et al. 2011, 2012</td>
<td>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.
3. branch-site model: “model B”

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

- Pairwise methods have very low power to detect adaptive evolution (via counting or likelihood method), but they are very fast.

- Branch models allow variation among branches but assume one \( \omega \) for all sites, and have low power to detect positive selection. Good for detecting shifts in average selection intensity.

- Site models allow variation among sites but assume selection pressure does not change among branches, and will have higher power if positive selection is long term (it comes in other flavors!)

- Branch-site++ models are very difficult to use, as they require more data and often have multiple sub-optimal peaks (caution with genome scans!)
Let's take a break

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

1. What is the intensity of selection on coral GFPs
2. Have there been episodes of positive selection during the evolution of colour diversity?
3. Are some sites in GFPs positively selected?
4. Which sites?
5. What happens to the colour when the amino acid is changes at these sites?

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
ML parameter estimation

\[ t, \kappa, \omega = \text{unknown values estimated by ML} \]

\[ \pi's = \text{empirical} \quad \text{[GY: F3x4 or F61 in Lab]} \]

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

Parameters: \( t \) and \( \omega \)
Gene: acetylcholine \( \alpha \) receptor

\( \ln L = -2399 \)
Task 2

How do we know that the estimate is significant?

Task 1. Parameter estimation (e.g., $\omega$) ✔

Task 2. Hypothesis testing

Task 3. Prediction / Site identification

testing nested hypotheses by using the LRT

\[ \ell_0 \text{ is the maximum log likelihood under } H_0 \text{ with parameters } \theta_0 \]

and

\[ \ell_1 \text{ is the maximum log likelihood under } H_1 \text{ with parameters } \theta_1 \]

Test statistic

\[ 2\Delta \ell = 2(\ell_0(\theta_0) - \ell_1(\theta_1)) \]

Degrees of freedom = difference in the number of parameters between the two models
Task 2

testing nested hypotheses by using the LRT

<table>
<thead>
<tr>
<th>Model</th>
<th>Code</th>
<th>NP</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-ratio</td>
<td>M0</td>
<td>1</td>
<td>$\omega$</td>
</tr>
<tr>
<td>Neutral</td>
<td>M1a</td>
<td>2</td>
<td>$p_0, \omega$</td>
</tr>
<tr>
<td>Selection</td>
<td>M2a</td>
<td>4</td>
<td>$p_0, p_1, \omega_0, \omega_2$</td>
</tr>
<tr>
<td>Discrete</td>
<td>M3</td>
<td>2k+1</td>
<td>$p_0, p_1, \ldots, p_k, \omega_0, \omega_1, \ldots, \omega_k$</td>
</tr>
<tr>
<td>Frequency</td>
<td>M4</td>
<td>5</td>
<td>$p_0, p_1, \ldots, p_4$</td>
</tr>
<tr>
<td>Gamma</td>
<td>M5</td>
<td>2</td>
<td>$\alpha, \beta$</td>
</tr>
<tr>
<td>2Gamma</td>
<td>M6</td>
<td>4</td>
<td>$p_0, \alpha_0, \beta_0, \alpha_1$</td>
</tr>
<tr>
<td>Beta</td>
<td>M7</td>
<td>2</td>
<td>$p, q$</td>
</tr>
<tr>
<td>Beta&amp;$\omega$</td>
<td>M8</td>
<td>4</td>
<td>$p_0, p, q, \omega$</td>
</tr>
<tr>
<td>Beta&amp;$\gamma$</td>
<td>M9</td>
<td>5</td>
<td>$p_0, p, q, \alpha, \beta$</td>
</tr>
<tr>
<td>Beta&amp;normal+1</td>
<td>M10</td>
<td>5</td>
<td>$p_0, p, q, \alpha, \beta$</td>
</tr>
<tr>
<td>Beta&amp;normal+1</td>
<td>M11</td>
<td>5</td>
<td>$p_0, p, q, \mu, \sigma$</td>
</tr>
<tr>
<td>0&amp;2normal+1</td>
<td>M12</td>
<td>5</td>
<td>$p_0, p_1, \mu_0, \sigma_1, \sigma_2$</td>
</tr>
<tr>
<td>3normal+0</td>
<td>M13</td>
<td>6</td>
<td>$p_0, p_1, \mu_0, \sigma_0, \sigma_1, \sigma_2$</td>
</tr>
</tbody>
</table>

Task 2

LRT No. 1: Does selection pressure vary among sites?

$H_0$: uniform selective pressure among sites (M0)
$H_1$: variable selective pressure among sites (M3)

Compare $2\Delta l = 2(l_1 - l_2)$ with a $\chi^2$ distribution

Model 0

\[ \hat{\omega} = 0.65 \]

Model 3

\[ \hat{\omega} = 0.01 \]

\[ \hat{\omega} = 0.90 \]

\[ \hat{\omega} = 5.55 \]
**Task 2**

LRT No. 2: Have some sites evolved under 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](image1.png) ![Model 2a](image2.png)

**Task 2**

LRT No. 3: Have some sites evolved under 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

![M7: beta](image3.png) ![M8: beta&\omega](image4.png)
the LRT does not follow the $\chi^2$ distribution

Data from: Anisimova, Bielawski, and Yang (2001) MBE 18: 1585-1592

Task 2

the LRT is conservative

Number of cases out of 100 for which the null hypothesis was rejected at the $\alpha = 1\%$ (5%) significance levels

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Simulation</th>
<th>LRT</th>
<th>Simulation parameters</th>
<th>Type I error at $\alpha = 1%$ (5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\kappa$ $\omega$ $\ell$ $f$ $N = 100$ $N = 500$</td>
<td></td>
</tr>
<tr>
<td>A...</td>
<td>M0</td>
<td>M0 &amp; M3</td>
<td>6 2 0.40</td>
<td>0.11 0 (0) 0 (0) 1.1 0 (0) 0 (0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11 0 (0) 0 (0)</td>
</tr>
<tr>
<td>B...</td>
<td>M0</td>
<td>M0 &amp; M3</td>
<td>17 2 0.40</td>
<td>2.11 0 (0) 0 (1) 8.44 0 (0) 0 (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16.88 0 (1) 0 (0)</td>
</tr>
<tr>
<td>C...</td>
<td>M0</td>
<td>M0 &amp; M3</td>
<td>5 5 0.25</td>
<td>0.91 0 (0) 0 (0) 9.1 0 (0) 0 (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18.2 0 (1) 2 (3)</td>
</tr>
<tr>
<td>D...</td>
<td>M7</td>
<td>M7 &amp; M8</td>
<td>6 2 $p = 0.41$ $q = 1.10$</td>
<td>0.11 N/A 0 (0) 1.1 N/A 1 (5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11 N/A 1 (4)</td>
</tr>
</tbody>
</table>

NOTE: Here $\ell$ denotes total tree length (sum of all branch lengths in the tree)

Data from: Anisimova, Bielawski, and Yang (2001) MBE 18: 1585-1592
the LRT is can be powerful

Power of the LRT: Number of replicates out of 100 in which positive selection was indicated by parameter estimates ($P_+$), or detected by the LRT at the 1% ($P_{+, 0.01}$) and 5% ($P_{+, 0.05}$, in parentheses) significance levels

<table>
<thead>
<tr>
<th>Simulation</th>
<th>LRT</th>
<th>Simulated taxa</th>
<th>$\kappa$</th>
<th>$\omega$ distribution</th>
<th>$I$</th>
<th>$L_C = 100$</th>
<th>$L_C = 500$</th>
<th>$L_C = 100$</th>
<th>$L_C = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M3</td>
<td>M0 &amp; M3</td>
<td>17</td>
<td>2</td>
<td>$\omega_0 = 0.018$, $p_0 = 0.366$; $\omega_1 = 0.304$, $p_1 = 0.532$; $\omega_2 = 1.691$, $p_2 = 0.079$</td>
<td>0.38</td>
<td>61</td>
<td>80</td>
<td>10 (17)</td>
<td>64 (72)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.11</td>
<td>93</td>
<td>100</td>
<td>91 (92)</td>
<td>100 (100)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.44</td>
<td>99</td>
<td>100</td>
<td>99 (99)</td>
<td>100 (100)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16.88</td>
<td>99</td>
<td>99</td>
<td>99 (99)</td>
<td>99 (99)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>105.5</td>
<td>31</td>
<td>58</td>
<td>31 (31)</td>
<td>58 (58)</td>
</tr>
</tbody>
</table>

NOTE: Here $I$ denotes total tree length (sum of all branch lengths in the tree)

Data from: Anisimova, Bielawski, and Yang (2001) MBE 18: 1585-1592

Task 3

How do we 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 $\omega > 1$?

**Model:**
5% have $\omega > 1$

**Bayes’ rule:**
site 4, 12 & 13

**Structure:**
sites are in contact

---

Bayes’ rule: yet another (silly) example of

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.

$Q_1$: What is the probability that any individual carries the disease?

$A_1$: $0.6 \times 0.01 + 0.4 \times 0.001 = 0.0064$

$P(D) = P(M)P(D|M) + P(F)P(D|F)$

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

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

See Yang and Bielawski (2000) TREE 15:496-503 for a detailed presentation of this example.

Task 3

identifying selected sites under a codon model

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

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

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

- 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

$P(\omega_2 | x_h) = \frac{p_2P(x_h | \omega_2)}{P(x_h)} = \frac{p_2P(x_h | \omega_2)}{\sum_{i=0}^{K-1} p_iP(x_h | \omega_i)}$

NOTE: The posterior probability should NOT be interpreted as a "P-value"; it can be interpreted as a measure of relative support.
Task 3

Empirical Bayes

Naive Empirical Bayes
• (NEB)
• Nielsen and Yang, 1998
• assumes no MLE errors

Bayes Empirical Bayes
• (BEB)
• Yang et al., 2005
• accommodate MLE errors

Task 3

Bayes Empirical Bayes (BEB)

1. Assign a prior to $\omega$ distribution parameters
2. Fix branch lengths to MLEs
3. Integrate over uncertainty
4. BEB is faster than “Full Bayes” (FB)

<table>
<thead>
<tr>
<th>False classification rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small datasets:</td>
</tr>
<tr>
<td>Large datasets:</td>
</tr>
</tbody>
</table>

* exception: extreme parameter estimates

Task 1. Parameter estimation (e.g., $\omega$)

Task 2. Hypothesis testing

Task 3. Prediction / Site identification

let's put this into practice ...

Is color diversity tuned by natural selection?
Is there a relationship between colour and endosymbiotic algae?
Questions we have:

1. What is the intensity of selection on coral GFPs
2. Have there been episodes of positive selection during the evolution of colour diversity?
3. Are some sites in GFPs positively selected?
4. Which sites?
5. What happens to the colour when the amino acid is changes at these sites?

Bayes’ rule:

\[ P(\omega_2 | x_h) = \frac{P(x_h | \omega_2) P(\omega_2)}{\sum \omega_i P(x_h | \omega_i)} \]

sites in red correspond to the protein-binding region of non-colored homologs of these GFP proteins.
The green fluorescent protein

signal 2: episodic selection

Baye’s rule:

\[
L(k_i | k_j) \propto \frac{P(k_i | k_j) P(k_j)}{P(k_i)}
\]

\(0.03\) \(0.59\) \(2.2\)

just for fun ....

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.
A model is an intentional simplification of a complex situation designed to eliminate extraneous detail in order to focus attention on the essentials of the situation” (Daniel L. Hartl)

“Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.” (George E. P. Box)

Sequence evolution is complex

How wrong is too wrong?

When your decisions about the details of your models lead to false biological conclusions !!!
sequence evolution is complex

1. codon usage
2. variation among sites
3. variation over time
4. recombination

Sums of codon usage counts the GstD gene of Drosophila

<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>2</td>
<td>0</td>
</tr>
<tr>
<td>TTC</td>
<td>2</td>
<td>55</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>TCA</td>
<td>0</td>
<td>15</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>TCG</td>
<td>1</td>
<td>15</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>TTA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TCA</td>
<td>0</td>
<td>15</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>TCG</td>
<td>1</td>
<td>15</td>
<td>4</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>5</td>
</tr>
<tr>
<td>ACT</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>GCA</td>
<td>2</td>
<td>7</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>GAA</td>
<td>2</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>GCG</td>
<td>3</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

how to model codon frequencies?
Biological conclusions are affected

Positive selection in highly biased gene sequences?

Estimation bias for the $d_N/d_S$ ratio
Simulation: GC3 = 89.5% (ENC = 28.3)

Unpublished simulation study

Biological conclusions are affected

What is the genomic relationship between $d_s$ and GC content in mammals?

Simple Model

Artiodactyla vs. Primates (82 nuclear genes)

how to model codon frequencies?

substitution rates are proportional to empirical frequency of:

<table>
<thead>
<tr>
<th>Model</th>
<th>TargetCodon</th>
<th>TargetNucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goldman and Yang 1994 (GY)</td>
<td>target codon</td>
<td>target nucleotide</td>
</tr>
<tr>
<td>Muse and Gaut 1994 (MG)</td>
<td>target codon</td>
<td>target nucleotide</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.

Example: A → C

AAA → CAA
AAA → ACA
AAA → AAC

Δ at codon position

<table>
<thead>
<tr>
<th></th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>GY</td>
<td>π_{CAA}</td>
<td>π_{ACA}</td>
<td>π_{AAC}</td>
</tr>
<tr>
<td>MG</td>
<td>π_c^1</td>
<td>π_c^2</td>
<td>π_c^3</td>
</tr>
</tbody>
</table>
Can choice of frequency model impact biological conclusions?


conditional nucleotide frequencies (CNF)

Example: A → C

AAA → CAA
AAA → ACA
AAA → AAC

<table>
<thead>
<tr>
<th>Δ at codon position</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>GY</td>
<td>π_{CAA}</td>
<td>π_{ACA}</td>
<td>π_{AAC}</td>
</tr>
<tr>
<td>MG</td>
<td>π_{c1}</td>
<td>π_{c2}</td>
<td>π_{c3}</td>
</tr>
<tr>
<td>CNF</td>
<td>π_{c1} \cdot π_{A,A} \cdot π_{A,A} \cdot π_{A,A}</td>
<td>π_{c2} \cdot π_{A,A} \cdot π_{A,A} \cdot π_{A,A}</td>
<td>π_{c3} \cdot π_{A,A} \cdot π_{A,A} \cdot π_{A,A}</td>
</tr>
</tbody>
</table>

how to model codon frequencies?

**Caution:** Inadequate, or inappropriate, modeling of codon usage can lead to estimation errors for the $\omega$ parameter!

- corrections for codon bias sometimes contain strong assumptions about the evolutionary process

- gene-level: errors can lead to false signal for $\omega > 1$

- genome-level: systematic errors can lead to false associations

Figure adapted from Yap et al. (2010) mbe 27: 726-734.
sequence evolution is complex

1. codon usage ✔
2. variation among sites
3. variation over time
4. Recombination

**M-Series models (and many others):** Biological interpretation of differences among sites in $\omega$ requires that such differences are due to selection pressure alone.

_Should we be concerned?_
transmembrane proteins

Partition sites within a gene:

- Loop structures extend into extra-cellular space: Hydrophilic amino acids here
  - \( \omega_0 \pi_0 K_0 C_0 \)
- Cell membrane in grey; helix structures span the membrane: Hydrophobic amino acids here
  - \( \omega_1 \pi_1 K_1 C_1 \)
- Loop structures extend into cytoplasm: Hydrophilic amino acids here
  - \( \omega_2 \pi_2 K_2 C_2 \)

GY-type codon models: variable \( \omega \) + c's among sites = variable \( d_s \) & \( d_a \) among sites

---

Realistic simulation w/ no positive selection

<table>
<thead>
<tr>
<th>error</th>
<th>site class 1</th>
<th>site class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2a*†</td>
<td>7.2% (7.7%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \omega_0 = 0.06; ) ( \pi_0 = 0.86; ) ( p_1 = 0.83; ) ( p_2 = 0.13 )</td>
<td>( \omega_0 = 2.48 ) ( p_1 = 0.04 )</td>
</tr>
<tr>
<td>M3†</td>
<td>6.4%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \omega_1 = 0.07; ) ( \pi_1 = 0.86 )</td>
<td>( \omega_0 = 1.45 ) ( p_1 = 0.14 )</td>
</tr>
<tr>
<td>M8*†</td>
<td>5.5% (6.0%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \omega_1 = 0.33; ) ( \pi_1 = 2.51 ) ( p_1 = 0.91 )</td>
<td>( \omega_0 = 1.92 ) ( p_1 = 0.09 )</td>
</tr>
<tr>
<td>Soft LiBaC 1</td>
<td>10.1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \omega_1 = 0.04; ) ( \pi_1 = 0.79 )</td>
<td>( \omega_0 = 0.72 ) ( p_1 = 0.21 )</td>
</tr>
<tr>
<td>Soft LiBaC 2</td>
<td>9.6%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \omega_1 = 0.05; ) ( \pi_1 = 0.80 )</td>
<td>( \omega_0 = 0.72 ) ( p_1 = 0.20 )</td>
</tr>
</tbody>
</table>

* LRTs for positive selection were highly significant in all cases.  
† High posterior probabilities for sites having \( \omega > 1 \) in all replications.
A sample of eight transmembrane proteins from *Rickettsia*. Results are for the group of sites having the largest value of $\omega$.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Nc</th>
<th>M2a</th>
<th>M8</th>
<th>LiBaC-1</th>
<th>LiBaC-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tbl-VirB6_3</td>
<td>938</td>
<td>$\omega = 5.8^*$</td>
<td>$\omega = 4.3^*$</td>
<td>$\omega = 0.43$</td>
<td>$\omega = 0.42$</td>
</tr>
<tr>
<td>Rta</td>
<td>403</td>
<td>$\omega = 4.29^*$</td>
<td>$\omega = 3.35^*$</td>
<td>$\omega = 1.73$</td>
<td>$\omega = 1.31$</td>
</tr>
<tr>
<td>Ccmf</td>
<td>635</td>
<td>$\omega = 15.5^*$</td>
<td>$\omega = 5.57^*$</td>
<td>$\omega = 1.07$</td>
<td>$\omega = 2.80$</td>
</tr>
<tr>
<td>nuoL3</td>
<td>499</td>
<td>$\omega = 12.5^*$</td>
<td>$\omega = 10.4^*$</td>
<td>$\omega = 1.45$</td>
<td>$\omega = 1.39$</td>
</tr>
<tr>
<td>Tbl-VirB6_2</td>
<td>657</td>
<td>$\omega = 32.8$</td>
<td>$\omega = 1.79^*$</td>
<td>$\omega = 0.44$</td>
<td>$\omega = 0.45$</td>
</tr>
<tr>
<td>perM</td>
<td>351</td>
<td>$\omega = 2.57$</td>
<td>$\omega = 2.91^*$</td>
<td>$\omega = 0.26$</td>
<td>$\omega = 0.10$</td>
</tr>
<tr>
<td>mivN</td>
<td>504</td>
<td>$\omega = 5.95$</td>
<td>$\omega = 2.52^*$</td>
<td>$\omega = 0.15$</td>
<td>$\omega = 0.18$</td>
</tr>
<tr>
<td>pgpA</td>
<td>198</td>
<td>$\omega = 35.0$</td>
<td>$\omega = 3.60^*$</td>
<td>$\omega = 0.57$</td>
<td>$\omega = 0.31$</td>
</tr>
</tbody>
</table>

**Caution**: Inadequate modeling of the complexity of among sites variation can lead to estimation biases.

- un-modeled site-variability can be “soaked-up” by other parameters free to vary among sites.
- if $\omega$ is the only parameter free to vary among sites, estimated differences among sites might not be due to selection pressure alone.
- this can sometime lead to **false signal for positive selection**
sequence evolution is complex

1. codon usage ✔
2. variation among sites ✔
3. variation over time
4. Recombination

The Cyanophage Molecular Mixing Bowl of Photosynthesis Genes

Synopses of Research Articles

“know your data”
pebS: evolution of a new phage gene

M3: $t = 28.8$ (LogDet: $t = 1.8$)

Phage: PebS GC3 = 30%

Prochlorococcus: PebA GC3 = 15%

"know your data"

HL and LL adapted Prochlorococcus ecotypes

M3: $t = 24.0$ (LogDet: $t = 1.5$)

HL adapted GC ≈ 29%

LL adapted GC ≈ 45%
Table: Highly biased branch-specific summary statistics under M0

<table>
<thead>
<tr>
<th>Branch</th>
<th>t</th>
<th>N</th>
<th>S</th>
<th>dN</th>
<th>dS</th>
<th>N*dN</th>
<th>S*dS</th>
</tr>
</thead>
<tbody>
<tr>
<td>18..19</td>
<td>6.118</td>
<td>375.3</td>
<td>134.7</td>
<td>0.5244</td>
<td>6.2595</td>
<td>196.8</td>
<td>843.4</td>
</tr>
<tr>
<td>F3x4</td>
<td>10.461</td>
<td>401.2</td>
<td>108.8</td>
<td>0.5446</td>
<td>14.3384</td>
<td>218.5</td>
<td>1559.9</td>
</tr>
<tr>
<td>F61</td>
<td>10.416</td>
<td>411.7</td>
<td>98.3</td>
<td>0.5504</td>
<td>15.7089</td>
<td>226.6</td>
<td>1544.2</td>
</tr>
<tr>
<td>9..15*</td>
<td>1.622</td>
<td>400.8</td>
<td>139.2</td>
<td>0.0648</td>
<td>1.9111</td>
<td>26</td>
<td>266</td>
</tr>
<tr>
<td>F3x4</td>
<td>8.07</td>
<td>409.1</td>
<td>130.9</td>
<td>0.0624</td>
<td>10.9047</td>
<td>25.5</td>
<td>1427.1</td>
</tr>
<tr>
<td>F61</td>
<td>15.343</td>
<td>413.2</td>
<td>126.8</td>
<td>0.0611</td>
<td>21.5828</td>
<td>25.3</td>
<td>2736.4</td>
</tr>
</tbody>
</table>

* Data for branch 9..15 obtained from a separate analysis of ONLY LL adapted Prochlorococcus

**know your data**

@!#%^! ... that’s a lot of substitutions

simulate under the null (no positive selection)

**know your data**
The ugly: non-stationary models

**Caution:** Serious estimation errors can arise when the process of evolution does not follow the same stochastic pattern in different lineages.

- Given a large shift, the average frequencies can sometimes be a poor fit to the data.
- Large changes in codon frequencies (non-homogeneity) can lead to very large errors in branch lengths.
- The combined effect of non-homogenous codon frequencies and shifts in the selection intensity can lead to false signal for positive selection.
- Parameter estimates, and other summary statistics, should be regularly inspected for problem signs.

Lower conditional likelihood (lcL)

Rooted conditional likelihood (rcL)

\[
L_{\text{root}} = \sum_{t} \left\{ \text{rcL} \times \sum_{x} P_{x,t}(t, x) \times \sum_{y} P_{y,t}(t, y) \times kL_{x}(x) \times \sum_{z} P_{z,t}(t, z) \sum_{w} P_{w,t}(t, w) \times kL_{z}(z) \right\}
\]
sequence evolution is complex

1. codon usage ✔
2. variation among sites ✔
3. variation over time ✔
4. Recombination

Note: Recombination adds among sites variability on both process and phylogeny!
Task 3

performance: model misspecification

**Caution:** high levels of recombination lead to type I errors

- Low recombination: LRT is robust (<15% of branches)
- High recombination: LRT type I error rate as high as 90%
- Bayesian site identification is less sensitive than LRT

For more details:

For a solution:

"**know your data**"

“There is no true interpretation of anything; interpretation is a vehicle in the service of human comprehension. The value of interpretation is in enabling others to fruitfully think about an idea”

—Andreas Buja
“We are drowning in information and starving for knowledge”

— Rutherford D. Roger