Codon substitution models and analysis of natural selection pressure in protein-coding genes

Joseph P. Bielawski

Department of Biology
Department of Mathematics & Statistics
Dalhousie University

Codon models have many uses:

Investigate process of protein evolution [this talk]
Phylogenetic inference
Ancestral sequence reconstruction
Dating divergence events
Alignment
Morphological adaptation:

Molecules

Troponin C: fast skeletal  Troponin C: cardiac and slow skeletal
Molecular Adaptation

“it has proven remarkably difficult to get compelling evidence for changes in enzymes brought about by selection, not to speak of adaptive changes…” (Lewontin, 1979).

Requirements:

1. DNA and protein sequences
2. Phylogeny
3. Phenotypes
4. Powerful analytical tools
Overview:

1. Brief overview (pop-gen & comparative data)
2. The $d_{W}/d_{s}$ ratio: the basics (and some complications)
3. Markov models of codon evolution
4. Three tasks and some guidelines

This morning

5. New performance issues

6. Accounts of selection

This afternoon

7. PAML introduction

8. Real data exercises

“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
**Single-locus tests:**

Composite null hypothesis: strict neutrality + constant/single population
Alternative hypothesis: selection and/or altered population demography

- Neutrality is typically harder to reject than demographic deviations
- Neutrality tests do not distinguish between different forms of selection

Genomic context is unknown
- Demography affects all loci in the genome, selection affects specific loci

**Multiple-locus tests:**

- Permit genomic approach to testing for selection
- Demographic processes might inflate variance of some test statistics

---

**A wide variety of approaches:**

<table>
<thead>
<tr>
<th>Sample data</th>
<th>Type of test</th>
<th>Data patterns</th>
<th>Family of tests</th>
<th>Robust to demography</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>Single-locus</td>
<td>Frequency spectrum</td>
<td>Ewans-Waterson test; tests related to Tajima's D-test</td>
<td>NO</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Linkage disequilibrium</td>
<td>Haplotype tests; test of the No. of alleles in a sample</td>
<td>NO</td>
</tr>
<tr>
<td>Multiple-loci</td>
<td>Population subdivision</td>
<td></td>
<td>$F_{st}$-related tests</td>
<td>?</td>
</tr>
<tr>
<td>Population + comparative data</td>
<td>Single-locus</td>
<td>Synonymous &amp; nonsynonymous polymorphisms</td>
<td>MacDonald-Kreitman type tests; PRF-tests</td>
<td>YES</td>
</tr>
<tr>
<td></td>
<td>Multiple-loci</td>
<td>Polymorphism/substitutions</td>
<td>HKA test</td>
<td>NO</td>
</tr>
</tbody>
</table>

A list of references will be posted on the course website.
1. **Counting methods for estimating natural selection pressure**
   - Many many different methods

2. **Markov models for estimating natural selection pressure**

\[
\frac{d_n}{d_S} = 1: \text{neutral evolution}
\]
\[
\frac{d_n}{d_S} < 1: \text{purifying (negative) selection}
\]
\[
\frac{d_n}{d_S} > 1: \text{diversifying (positive) selection}
\]

Kimura (1968)

---

**Proteins**: Fast evolving genes or deep divergences lead to saturation of \(d_S\).

**Classify sites according to co-evolution of amino acids**:

**Classify sites according to rate shifts along lineages or in clades**
- e.g., Miyamoto and Fitch 1995; Gaucher, et al. 2001; Knudsen & Miyamoto, 2001; Gu, 2001; Pupko and Galtier, 2002; Siltberg and Liberales 2002; Blouin et al. 2003; Knudesen et al. 2003

**Classify sites according shifts in physiochemical properties**
Keynotes

- We have a wide variety of approaches to choose from
- Very large system data (genome-scale approaches)
- Methods robust to demography are being developed
- Comparative results confirmed with population data
- New methods permit different indices of selection pressure
- The variety of methods compliment each other very well

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

The basics

Hereafter we focus on codon-based methods
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.

Two types of changes among codons:

- **synonymous**: TTT (Phe) → TTC (Phe)
- **nonsynonymous**: TTT (Phe) → TTA (Leu)
**Define two rates**

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

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

The \( \omega = d_N/d_S \) ratio measures selection at the protein level

\[ \omega = 1: \text{neutral evolution} \]
\[ \omega < 1: \text{purifying (negative) selection} \]
\[ \omega > 1: \text{diversifying (positive) selection} \]

Kimura (1968)
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>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All 3 Positions</td>
</tr>
<tr>
<td>Total mutations</td>
<td>549 (100)</td>
</tr>
<tr>
<td>Synonymous</td>
<td>134 (25)</td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>392 (71)</td>
</tr>
<tr>
<td>nonsense</td>
<td>23 (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.
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, $d_s = \frac{5}{229.5} = 0.0218$
\[ d_N = \frac{5}{670.5} = 0.0075 \]

$d_n/d_s (\omega) = 0.34$, **purifying selection !!!**

We start with pairwise counting methods:

1. Simple
2. Intuitive
3. Popular
Estimation of $d_S$ and $d_N$ between two sequences (pairwise) is a very popular approach.

<table>
<thead>
<tr>
<th></th>
<th>human</th>
<th>Cow</th>
<th>Rabbit</th>
<th>Rat</th>
<th>Mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GTG</td>
<td>CTG</td>
<td>CCT</td>
<td>GCC</td>
<td>GAC</td>
</tr>
<tr>
<td>human</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cow</td>
<td>...</td>
<td>...</td>
<td>G.C</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Rabbit</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>A.</td>
</tr>
<tr>
<td>Rat</td>
<td>...</td>
<td>C</td>
<td>.AT</td>
<td>A</td>
<td>...</td>
</tr>
<tr>
<td>Mouse</td>
<td>...</td>
<td>C</td>
<td>GA</td>
<td>G</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>G</td>
</tr>
</tbody>
</table>

Counting methods

1. Count synonymous ($S$) and nonsynonymous ($N$) sites
2. Count synonymous and nonsynonymous differences
3. Apply some correction (e.g., correct for multiple hits)
Complications

Counting methods:


Step 1: Counting sites (*S* and *N*)

Codon TTT has:
- 9 neighbour codons
  - 1 synonymous neighbour
  - 8 non-synonymous neighbours

Frequency of mutational opportunities:
- 1/9 (11%) synonymous
- 8/9 (89%) nonsynonymous

*S* and *N*:
- \( S = 0.11 \times 3 = 0.33 \)
- \( N = 0.89 \times 3 = 2.67 \)
Counts of $S$ and $N$ are very sensitive to assumptions about the process of molecular evolution:

- $T_s / T_v$ ratio ($\kappa$)

- Nucleotide (codon) bias ($\pi$)

### Complications

Part 2: The $d_N / d_S$ ratio

Counting the number of synonymous and nonsynonymous sites

#### When transitions $=$ transversions ($\kappa = 1$)

<table>
<thead>
<tr>
<th>Target Codon</th>
<th>Type of change</th>
<th>Rate ($\kappa = 1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTC</td>
<td>synonymous</td>
<td>1</td>
</tr>
<tr>
<td>TTA</td>
<td>nonsynonymous</td>
<td>1</td>
</tr>
<tr>
<td>TTG</td>
<td>nonsynonymous</td>
<td>1</td>
</tr>
<tr>
<td>TCT</td>
<td>nonsynonymous</td>
<td>1</td>
</tr>
<tr>
<td>TAT</td>
<td>nonsynonymous</td>
<td>1</td>
</tr>
<tr>
<td>TGT</td>
<td>nonsynonymous</td>
<td>1</td>
</tr>
<tr>
<td>GCT</td>
<td>nonsynonymous</td>
<td>1</td>
</tr>
<tr>
<td>ATT</td>
<td>nonsynonymous</td>
<td>1</td>
</tr>
<tr>
<td>GTT</td>
<td>nonsynonymous</td>
<td>1</td>
</tr>
</tbody>
</table>

**Total = 9**

- $S$: $1/9 \times 3 = 1/3$ (0.33)
- $N$: $8/9 \times 3 = 8/3$ (2.67)

#### When transitions $=$ 2x transversions ($\kappa = 2$)

<table>
<thead>
<tr>
<th>Target Codon</th>
<th>Type of change</th>
<th>Rate ($\kappa = 2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTC</td>
<td>synonymous</td>
<td>2</td>
</tr>
<tr>
<td>TTA</td>
<td>nonsynonymous</td>
<td>1</td>
</tr>
<tr>
<td>TTG</td>
<td>nonsynonymous</td>
<td>1</td>
</tr>
<tr>
<td>TCT</td>
<td>nonsynonymous</td>
<td>2</td>
</tr>
<tr>
<td>TAT</td>
<td>nonsynonymous</td>
<td>1</td>
</tr>
<tr>
<td>TGT</td>
<td>nonsynonymous</td>
<td>1</td>
</tr>
<tr>
<td>GCT</td>
<td>nonsynonymous</td>
<td>2</td>
</tr>
<tr>
<td>ATT</td>
<td>nonsynonymous</td>
<td>1</td>
</tr>
<tr>
<td>GTT</td>
<td>nonsynonymous</td>
<td>1</td>
</tr>
</tbody>
</table>

**Total = 12**

- $S$: $2/12 \times 3 = 1/2$ (0.5)
- $N$: $10/12 \times 3 = 5/2$ (2.5)
At the third position, transitions are more likely to be synonymous than transversions.

Complications

Part 2: The \( \frac{d_\text{N}}{d_\text{S}} \) ratio

**Effect of assumptions on \( S \)**

![Graph showing effect of assumptions on \( S \)]

Data from: Bielawski, Dunn, and Yang (2000) Genetics, 156: 1299-1308

---

**Bias is complicated**

![Graph showing bias in method of Nei and Gojobori (1986)]

Data from: Bielawski, Dunn, and Yang (2000) Genetics, 156: 1299-1308
Step 2: Counting differences

Example: two pathways between CCT and CAG:

<table>
<thead>
<tr>
<th>Pathways</th>
<th>Syn</th>
<th>Nonsyn</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCT (Pro) ↔ CAT (His) ↔ CAG (Gln)</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>CCT (Pro) ↔ CCG (Pro) ↔ CAG (Gln)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>0.5</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Nearly all counting methods make the *ad hoc* assumption that all pathways are equally likely.
Step 3: Correcting for multiple hits

Corrections for multiple hits have been based on nucleotide substitution models. Nucleotide based corrections are **INVALID**, but errors will be low if sequence divergence is low.

Simulation study of estimation bias in $d_s$

Data from: Dunn, Bielawski, and Yang (2001) Genetics, 157: 295-305
Why are estimation biases worth worrying about?

Because these biases can lead to qualitatively different biological conclusions !!!

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

1. [Diagram showing a curve relationship]
2. [Diagram showing a straight line relationship]
3. [Diagram showing a linear increase relationship]

Most studies
Miata et al. 1989
Bernardi et al. 1993
Matassi et al. 1999

Eyre Walker 1994
Mammalian nuclear genes:

**Simple Model**

\[ r^2 = 0.0228, P = 0.1759 \]

**Model with ts/tv and codon bias**

\[ r^2 = 0.53, P < 0.0001 \]

Artiodactyla vs. Primates (82 nuclear genes)


---

**Estimation bias for the \( d_N/d_S \) ratio**

Simulation: GC3 = 89.5\% (ENC = 28.3)

[Graph showing relationship between sequence divergence and \( d_N/d_S \) ratio for positive and purifying selection.]
Complications

How to model codon bias?

Substitution rates are proportional to empirical frequency of:

Goldman and Yang 1994 (GY): target codon
Muse and Gaut 1994 (MG): target nucleotide

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.

Complications: how to model codon bias

Example: A → C
AAA → CAA
AAA → ACA
AAA → AAC

<table>
<thead>
<tr>
<th></th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>GY</td>
<td>$\pi_{CAA}$</td>
<td>$\pi_{ACA}$</td>
<td>$\pi_{AAC}$</td>
</tr>
<tr>
<td>MG</td>
<td>$\pi_c^1$</td>
<td>$\pi_c^2$</td>
<td>$\pi_c^3$</td>
</tr>
</tbody>
</table>
Two issues:

1. Detecting selection (we will return to this later)

2. Comparing rates of evolution among genes

Complications: how to model codon bias

Part 2: The $d_N/d_S$ ratio

The $d_{N}/d_{S}$ ($\omega$) ratio is a valuable index of selection pressure!

Computing the $d_{N}/d_{S}$ ($\omega$) ratio can be tricky!

**Keynotes**

1. **Assumptions matter more than methods** (usually).

2. Ignoring the transition/transversion rate bias leads to underestimation of $S$, overestimation of $d_{S}$ and underestimation of the $d_{N}/d_{S}$ ($\omega$) ratio.

3. Codon-usage bias often has the opposite effect to the transition/transversion bias and can be more important.

4. Different assumptions can produce different estimates even when the sequences are highly similar.

5. Different assumptions can lead to different biological conclusions (**You have been warned!**
Markov models of codon evolution

Goldman & Yang 1994 *MBE* **11**:725-736

Muse & Gaut 1994 *MBE* **11**:715-724

Why use a likelihood model of codon evolution?

1. We can take advantage of the phylogeny

2. Computation of transition probabilities accomplishes the following in 1 step:
   i. estimation of parameters \((t, \kappa, \omega)\)
   ii. correction for multiple hits
   iii. weight evolutionary pathways between codons
“GY” Codon models

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

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

Nonsynonymous

GTG (Val) $\rightarrow$ CTG (Leu): $\omega\pi_{CTG}$
CCG (Pro) $\rightarrow$ CTG (Leu): $\kappa\omega\pi_{CTG}$
Part 3: Codon models

The "GY" model

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

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

\[ P(t) = \{p_{ij}(t)\} = e^{qt} \]

The "MG" model

(Muse & Gaut 1994 MBE 11:715-724)

\[ q_{ij} = \begin{cases} 
0 & \text{if } i \text{ and } j \text{ differ at 2 or 3 positions} \\
\alpha\pi_n^m, & \text{for syn. transversion} \\
\alpha\kappa\pi_n^m, & \text{for syn. transition} \\
\beta\pi_n^m, & \text{for nonsyn. transversion} \\
\beta\kappa\pi_n^m, & \text{for nonsyn. transition} 
\end{cases} \]

\[ P(t) = \{p_{ij}(t)\} = e^{qt} \]

Note that \( \kappa \) was not included in the original model; some versions use "R", the tv/ts ratio, rather than \( r \) (ts/tv).
Simple case (only two codons):

\[
\begin{array}{c}
\text{CCC} \\
t_0 \\
k \\
\text{CCT} \\
t_1
\end{array}
\]

\[
L_h(\text{CCC}, \text{CCT}) = \sum_k \pi_k p_{\text{CCC}}(t_0)p_{\text{CCT}}(t_1)
\]

Note: analysis is typically done by using an unrooted tree

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.

\[
\ell(t, \kappa, \omega) = \ln\{L\} = \ln\{L_1\} + \ln\{L_2\} + \ln\{L_3\} + \ldots + \ln\{L_N\} = \sum_{h=1}^{N} \ln\{L_h\}
\]
Remember: we are interested in adaptive evolution

\( \omega = 1 \): neutral evolution

\( \omega < 1 \): purifying (negative) selection

\( \omega > 1 \): diversifying (positive) selection

**Problem: averaging over a pair**

In a pairwise analysis we must average the \( \omega \) ratio over:

1. all sites
2. the entire evolutionary history

In a large-scale pairwise database search, only 17 out of 3,595 genes were found to be under positive selection, at \(<0.5\%\) (Endo et al. 1996 *MBE* 13: 685-690)
Our question: **When?**

<table>
<thead>
<tr>
<th>b.l. (my)</th>
<th>Fraction of t</th>
<th>$\omega$</th>
</tr>
</thead>
<tbody>
<tr>
<td>115</td>
<td>0.203</td>
<td>0.2</td>
</tr>
<tr>
<td>55</td>
<td>0.097</td>
<td>0.2</td>
</tr>
<tr>
<td>60</td>
<td>0.106</td>
<td>0.2</td>
</tr>
<tr>
<td>60</td>
<td>0.106</td>
<td>0.2</td>
</tr>
<tr>
<td>85</td>
<td>0.150</td>
<td>0.5</td>
</tr>
<tr>
<td>35</td>
<td>0.062</td>
<td>0.5</td>
</tr>
<tr>
<td>35</td>
<td>0.062</td>
<td>0.5</td>
</tr>
<tr>
<td>120</td>
<td>0.212</td>
<td>1.2</td>
</tr>
</tbody>
</table>

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?

Selection pressure varies among sites

ATG CTG GIG CTA .............. .............. .............. .............. .............. .............. .............. CGC TAA

If we average over sites, we do NOT detect positive selection; \( \omega = 0.31 \)
**Problem:** averaging over a pair has very low power if the questions are about "when" or "where"!

**Solution:** Phylogenetic estimation of selection pressure

- variable $\omega$ over branches (when?)
- variable $\omega$ over sites (where?)
- variable $\omega$ over branches and sites (when and where?)

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

$$L(x_k) = \sum_k \sum_j \left[ \pi_k p_{kx_4}(t_4) p_{kx_3}(t_3) p_{kx_2}(t_0) p_{jx_2}(t_2) p_{jx_1}(t_1) \right]$$
\[ L(x_{i,j}) = \sum_k \sum_j \tau_k \rho_{kx_1} (t_4; \omega_0) \rho_{kx_3} (t_3; \omega_0) \rho_{t_0} (t_0; \omega_0) \rho_{jx_2} (t_2; \omega_0) \rho_{jx_1} (t_1; \omega_0) \]

Yang, 1998: fixed effects
Bielawski and Yang, 2003: fixed effects
Seo et al. 2004: auto-correlated rates
Kosakovsky Pond and Frost, 2005: genetic algorithm
Variation among sites:

Yang and Swanson, 2002:
Bao, Gu and Bielawski, 2006:
Suzuki and Gojobori, 1999;
Massingham and Goldman, 2005:
Kosakovsky Pond and Frost, 2005:
Nielsen and Yang, 1998:
Kosakovsky Pond, Frost and Muse, 2005:
Huelsenbeck and Dyer, 2004:
Huelsenbeck et al. 2006:
Bao, Gu, Dunn and Bielawski 2008:

Approach:
fixed effects
fixed effects
site wise, counts
site wise, LRT
site wise, LRT
random effects
random effects
full Bayesian
full Bayesian
LiBaC (MBC)

* This is not a comprehensive list

Site models

- Vaccine design
- Genetic incompatibilities in human infertility
- Non-hormonal contraception drugs
- Identify pathogenicity genes
- Identify candidate genes for drug therapies
- Identify immune and defense system genes
- Aid functional classification of unknown genes
- Incorporate in models of protein 2D and 3D structure
Site models (over a phylogeny):
1. Counting methods
2. Fixed effect codon models
3. Random effect codon models ("M-series" models*)
4. LiBaC (Hard & Soft Model Based Clustering)

* we will review a very select set of examples
Random-effect site model "M3"

Discrete model (M3) with \( K = 3 \) site classes

- Site class 1: \( \omega_0 = 0 \), 60% of codon sites
- Site class 2: \( \omega_1 = 1 \), 38% of codon sites
- Site class 3: \( \omega_2 = 5 \), 02% of codon sites

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

"M-series" models for variable \( \omega \) among sites

<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, 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-2}, \alpha_0, \ldots, \alpha_{K-2} )</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, 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&gt;1</td>
<td>M11</td>
<td>5</td>
<td>( p_0, p, q, \mu, \sigma )</td>
</tr>
<tr>
<td>0&amp;2normal&gt;1</td>
<td>M12</td>
<td>5</td>
<td>( p_0, p_1, \mu_0, \sigma_0, \sigma_2 )</td>
</tr>
<tr>
<td>3normal&gt;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>
Yang and Nielsen, 2002: ML
Forsberg and Christiansen, 2003: ML
Bielawski and Yang, 2004: ML
Giundon et al., 2004: full Bayesian
Zhang et al. 2005: ML

Branch-site models

Site model: M3
Branch-site model: Model-B

$\omega = 0.01$  $\omega = 0.90$  $\omega = 5.55$

$\omega$ for background branches are from site-classes 1 and 2 (0.01 or 0.90)
1. Pairwise methods have very low power to detect adaptive evolution.

2. Branch models allow variation among branches but assume one $\omega$ for all sites, and have low power to detect positive selection.

3. 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!)

4. Branch-site models are very difficult to use, as they require more data and often have multiple sub-optimal peaks (caution with genome scans!)

“\textit{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).
Three tasks

1. Parameter estimation
2. Hypothesis testing
3. Site identification

Task 1: ML parameter \((t, \kappa, \omega)\) estimation:

\(t, \kappa, \omega\) = unknown values

\(\pi's\) = empirical [GY: F3\times4 or F61]

Use numerical hill-climbing algorithm to maximize the likelihood function
Task 1: Estimation of $d_s$ and $d_n$

- Numbers of substitutions are calculated from $q_{ij}$ and $t$.

- Number of sites ($S$ and $N$) are calculated from $q_{ij}$ by fixing $\omega = 1$.

**Software to estimate \( d_s \) and \( d_N \)**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting methods</td>
<td>MEGA; codeml &amp; yn00 in PAML</td>
</tr>
<tr>
<td>NG86</td>
<td></td>
</tr>
<tr>
<td>Li93</td>
<td>DAMBE</td>
</tr>
<tr>
<td>Comeron 95</td>
<td>DIVERGE by Comeron</td>
</tr>
<tr>
<td>YN00</td>
<td>yn00 in PAML</td>
</tr>
<tr>
<td>ML methods</td>
<td>codeml in PAML</td>
</tr>
<tr>
<td>GY94</td>
<td></td>
</tr>
<tr>
<td>MG94</td>
<td>HYPHY</td>
</tr>
</tbody>
</table>

* This is not a comprehensive list

---

**Task 2: hypothesis testing by using the LRT**

\[ t_0 \text{ is the maximum log likelihood under } H_0 \text{ with parameters } \theta_0 \]
and

\[ t_1 \text{ is the maximum log likelihood under } H_1 \text{ with parameters } \theta_1 \]

\[ \text{Test statistic } = 2\Delta \ell = 2(t_0(\theta_0) - t_1(\theta_1)) \]

Degrees of freedom = difference in the number of parameters between The two models
**Part 4: Three tasks**

**Task 2: hypothesis testing by using the LRT**

**H0**: uniform selective pressure among sites (M0)
**H1**: variable selective pressure among sites (M3)

Compare $2\Delta l = 2(l_1 - l_0)$ 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: hypothesis testing by using the LRT**

**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**

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

**Model 2a**

$\hat{\omega} = 0.5$  $(\omega = 1)$  $\hat{\omega} = 3.25$
Part 4: Three tasks

Task 2: hypothesis testing by using the LRT

**H$_0$**: Beta distributed variable selective pressure (M7)
**H$_1$**: Beta plus positive selection (M8)

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

**M7: beta**

**M8: beta&$\omega$**

Task 2: hypothesis testing by using the LRT

The LRT does not follow the $\chi^2$ distribution

Data from: Anisimova, Bielawski, and Yang (2001) MBE 18: 1585-1592
### Task 2: hypothesis testing by using the LRT

#### Part 4: Three tasks

**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></td>
<td>N = 100</td>
</tr>
<tr>
<td>A...</td>
<td>M0</td>
<td>M0 &amp; M3</td>
<td>$T\kappa \omega$ SN = 100</td>
<td>6 2 0.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B...</td>
<td>M0</td>
<td>M0 &amp; M3</td>
<td>$T\kappa \omega$ SN = 100</td>
<td>17 2 0.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C...</td>
<td>M0</td>
<td>M0 &amp; M3</td>
<td>$T\kappa \omega$ SN = 100</td>
<td>5 5 0.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D...</td>
<td>M7</td>
<td>M7 &amp; M8</td>
<td>$T\kappa \omega$ SN = 100</td>
<td>6 2 $p = 0.41$ $q = 1.10$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

...The LRT is conservative.

### Task 2: hypothesis testing by using the LRT

#### Part 4: Three tasks

**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% ($P_{+s, 0.01}$) and 5% ($P_{+s, 0.05}$ in parentheses) significance levels**

<table>
<thead>
<tr>
<th>Simulation</th>
<th>LRT</th>
<th>Simulation parameters</th>
<th>$P_+$</th>
<th>$P_{+s, 0.01}$ ($P_{+s, 0.05}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N3</td>
<td>M0 &amp; M3</td>
<td>$T\kappa \omega$ SN = 100</td>
<td>0.38 81 80</td>
<td>10 (17) 68 (72)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.11 93 100</td>
<td>91 (92) 100 (100)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8.44 99 100</td>
<td>99 (99) 100 (100)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16.88 99 99</td>
<td>99 (99) 99 (99)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>105.5 31 58</td>
<td>31 (31) 58 (58)</td>
</tr>
</tbody>
</table>

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

...The LRT can be powerful.
LRT can indicate the presence of sites evolving by positive selection

Parameter estimates tell us the fraction of such sites

How do we identify such sites?

1. "Site-wise methods"
2. Empirical Bayes
3. Hierarchical Bayes

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.

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

A₁: \(0.6 \times 0.01 + 0.4 \times 0.001 = 0.0064\)

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

See Yang and Bielawski (2000) TREE 15:496-503 for a detailed presentation of this example
Task 3: Example of Bayes’ rule

Q2: Given that an individual carries the disease, what is the probability that it is a male?

\[ A_2: 0.6 \times \frac{0.01}{0.0064} = 0.94 \]

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

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

Part 4: Three tasks

Discrete model (M3) with \( K = 3 \) site classes

- Site class 1: \( \omega_0 = 0 \), 60% of codon sites
- Site class 2: \( \omega_1 = 1 \), 38% of codon sites
- Site class 3: \( \omega_2 = 5 \), 02% of codon sites

\[
P(x_h) = \sum_{i=0}^{K-1} p_i \times P(x_h | \omega_i)
\]
Task 3: Bayes’ rule for identifying selected sites

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

Part 4: Three tasks

Empirical Bayes identification of positively selected sites

NOTE: The posterior probability should NOT be interpreted as a “P-value”; it can be interpreted as a measure of relative support
Task 3: an application

Bayesian site identification for **Green** absorbing PRs (branch-site model)

Task 3: an application

Bayesian site identification for **Blue** absorbing PRs (branch-site model)
**Task 3: Performance of site identification**

**Accuracy (precision):** the percentage of true positive selection sites among those classified as such.

**Power (recall)*:** the percentage of all the true positive selection sites in the data that are classified as such.

**Missclassification rate:** the total percentage of sites incorrectly misplaced from one site-class to another.

*In a shameless co-option of statistical terminology we used “power” to describe what is more formally referred to a “recall” or the “sensitivity.”

---

**Task 3: Site identification**

**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
Taxa = 17
ω = 2.08,
L_c = 500,
model = M8


Performance: NEB

Accuracy can be low for small datasets

Bayes Empirical Bayes

1. Assign a prior to $\omega$ distribution parameters
2. Fix branch lengths to MLEs
5. Integrate over uncertainty
6. BEB is faster than “Full Bayes”

<table>
<thead>
<tr>
<th>False positive rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small datasets:</td>
</tr>
<tr>
<td>BEB &lt; NEB</td>
</tr>
<tr>
<td>Large datasets:</td>
</tr>
<tr>
<td>BEB ≈ NEB*</td>
</tr>
</tbody>
</table>

* exception: extreme parameter estimates


Full (hierarchical) Bayes

- Full Bayes has minor advantages over BEB in small (test) data sets
- Full Bayes has advantages (NEB) when sequence divergence is low
- Full Bayes methods seem robust to light model misspecifications

Performance: model misspecification

**Warning:** Full Bayesian and Empirical Bayesian methods can be extremely sensitive to “heavy” model misspecification.

- Heavy model misspecification could lead to excessive type-I error rates
- More work is required to understand relationship between under-fitting/over-fitting analytical models and the generalization error rates (this is true for all approaches: site-wise, NEB, BEB and full Bayes)

**For more details see:**

Performance: recombinant data

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

- Low recombination: LRT is robust
- 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:**
Follow-up with experimental assessment of function

Part 4: Three tasks

1. **Computational analysis**

   Chromatogram (42) with 8 > 3 s/d columns

   \[ P(\omega_0 | x_0) = \frac{p_1 P(x_0 | \omega_0)}{P(x_0)} = \sum_{\omega} p_1 P(x_0 | \omega) \]

2. **Site identification and mapping**

3. **Reconstruct ancestral proteins**

   Site-directed mutagenesis

   Keynotes

1. LRT is accurate and powerful, even in small data sets
2. Parameter estimation is more difficult, and can be very sensitive to assumptions
3. Bayes site prediction is **very difficult** and requires much more data than LRT and parameter estimation
   1. Site prediction is NOT reliable from a few, similar sequences
   2. Site prediction can be sensitive to assumed \( \omega \) dist.
4. There is an optimal window of sequence divergence
5. Adding lineages is the most efficient way to improve power and accuracy
6. We recommend **robustness analyses**:  
   1. Use multiple models (M0-M3, M1a-M2a, M7-M8)  
   2. Evaluate sensitivity to model parameters (incl. tree!)  
   3. Attempt to identify "consensus sets of sites"
7. Check for and discard results from local optima
**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?*

---

**Some real datasets: transmembrane proteins**

Partition sites within a gene:

- Loop structures extend into extra-cellular space: $\omega_0 \pi_0 K_0 C_0$
  - Hydrophilic amino acids here
- 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 $d_0$ & $d_2$ among sites
Some real datasets: transmembrane proteins

Fixed-effect models

Model selection:
- AIC
- BIC
- Backward elimination (LRT)

Model selection: Backward elimination via LRT under $p = 0.001$; model of codon frequencies is F3x4; * Abolone sperm lysin is included because it is a case-study gene which has been analyzed under a wider variety of alternative models.

- Significant evolutionary heterogeneity among partitions
- Parameter estimates impacted by choice of model
Likelihood-based Clustering (LiBaC) is a method for grouping sites according to similarities in the underlying process of evolution.

FE models

LiBaC models

Variability among \( G \) groups of sites

<table>
<thead>
<tr>
<th>models</th>
<th>( \omega )</th>
<th>( \kappa )</th>
<th>( \zeta )</th>
<th>( \pi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiBaC-1</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>LiBaC-2</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
</tbody>
</table>

Note: Assumptions are expected to matter more than the method!

LiBaC is a fast, EM algorithm

- extends “model based clustering” (Fraley and Raferty, 1998)
- addresses the “classification problem”
- built upon the “FE models” (LiBaC1 from FE1, etc.)

\[
I(\theta, \theta^\text{old}) = \sum_{i=1}^{n} \sum_{k=1}^{G} P(z_{ik} = 1 | x, \theta^\text{old}) \log \left[ \tau_{\theta^\text{old}}(x_i | \theta_k) \right]
\]

LiBaC evaluates the posterior distribution of the site assignments and then estimates the joint probability of the complete data under the posterior distribution.

**Part 5: LiBaC on codon models**

"Soft-LiBaC"

**Initial step:** Obtain "initials" from a mixture model such as M3 or based on structure.

**E-step:** Based on the current parameter estimates, $\theta$, compute the posterior probabilities by Bayes' rule:

$$P(z_a = 1 | x, \theta) = \frac{r_{z_a} f(x_i | \theta_j)}{\sum_{j=1}^{M} r_j f(x_i | \theta_j)}$$

**M-step:** Re-estimate parameters $\theta = (\tau, \pi, \kappa, \omega)$ by maximizing $l(\theta, \theta^m)$ using the current posterior probabilities:

$$w_i = \frac{\sum_{z=1}^{M} P(z = 1 | x, \theta)}{\sum_{z=1}^{M} \sum_{i=1}^{n} P(z = 1 | x, \theta) \log [r_{z_a} f(x_i | \theta_j)]}$$

**Convergence:** Check for convergence of the log likelihood $l(\theta, \theta^m)$ evaluated at $\theta = \theta^m$. If the convergence criterion is not satisfied return to **E-step**.

---


---

**Part 5: New performance issues**

"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
Introducing the **Bayes error rate**

Bayes error rate: The expected lower boundary on classification, with the expectation taken on the true model and parameters. (e.g., Tumer and Ghosh. 2003. Int J Smart Eng Syst Des. 5:95-110.)

**Dataset 1**

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega$</td>
<td>0.3</td>
<td>1.5</td>
</tr>
<tr>
<td>$c$</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>$\pi$</td>
<td>1/61</td>
<td>empirical</td>
</tr>
</tbody>
</table>

Bayes error rate: 8.69%

**Dataset 2**

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega$</td>
<td>0.3</td>
<td>1.5</td>
</tr>
<tr>
<td>$c$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$\pi$</td>
<td>1/61</td>
<td>1/61</td>
</tr>
</tbody>
</table>

Bayes error rate: 26.8%

**Classification of sites having $\omega > 1$**

"the easy dataset"

"the hard dataset"

The idealized precision-recall curve gives the theoretical upper-bound on precision for each value of recall.
A hypothetical example

Method 1
Post cut off: 50%
Precision: 100%
Recall: 0%

Method 2
Post cut off: 50%
Precision: 82%
Recall: 20%

Method 3
Post cut off: 50%
Precision: 75%
Recall: 62%

Part 5: The Bayes error rate

M8 (BEB) 0.9
M3 (NEB) 0.9

Now, back to "datasets 1 and 2"

"the easy dataset"

"the hard dataset"
more simulation studies (all have a group with $\omega > 1$)

Some real transmembrane proteins: Uh-Oh!

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>LiBaC-1</th>
<th>LiBaC-2</th>
<th>M2a</th>
<th>M8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trbl-VirB6_3</td>
<td>938</td>
<td>0.43</td>
<td>0.42</td>
<td>5.8*</td>
<td>4.3*</td>
</tr>
<tr>
<td>Rfai</td>
<td>403</td>
<td>1.73</td>
<td>1.31</td>
<td>4.29*</td>
<td>3.35*</td>
</tr>
<tr>
<td>Comf</td>
<td>635</td>
<td>1.07</td>
<td>2.80</td>
<td>15.5*</td>
<td>5.57*</td>
</tr>
<tr>
<td>nul3</td>
<td>499</td>
<td>1.45</td>
<td>1.39</td>
<td>12.5*</td>
<td>10.4*</td>
</tr>
<tr>
<td>Trbl-VirB6_2</td>
<td>657</td>
<td>0.44</td>
<td>0.45</td>
<td>32.8</td>
<td>1.79*</td>
</tr>
<tr>
<td>perM</td>
<td>351</td>
<td>0.26</td>
<td>0.10</td>
<td>2.57</td>
<td>2.91*</td>
</tr>
<tr>
<td>mivN</td>
<td>504</td>
<td>0.15</td>
<td>0.18</td>
<td>5.95</td>
<td>2.52*</td>
</tr>
<tr>
<td>pgpA</td>
<td>198</td>
<td>0.57</td>
<td>0.31</td>
<td>35.0</td>
<td>3.60*</td>
</tr>
</tbody>
</table>

We need another simulation study: cases where there is NO positive selection.
Part 5: New performance issues

Another simulation: no positive selection this time!

<table>
<thead>
<tr>
<th>error</th>
<th>site class 1</th>
<th>site class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft LiBaC 1</td>
<td>10.1%</td>
<td>$\omega_1 = 0.04$ ($p_1 = 0.79$)</td>
</tr>
<tr>
<td>Soft LiBaC 2</td>
<td>9.6%</td>
<td>$\omega_1 = 0.05$ ($p_1 = 0.86$)</td>
</tr>
<tr>
<td>M2a†</td>
<td>7.2% (7.7%)</td>
<td>$\omega_1 = 0.06$; $\omega_2 = 2.48$</td>
</tr>
<tr>
<td>M3†</td>
<td>6.4%</td>
<td>$\omega_1 = 0.07$; $\omega_2 = 1.45$</td>
</tr>
<tr>
<td>MB*†</td>
<td>5.5% (6.0%)</td>
<td>$p = 0.33$; $p_1 = 0.91$</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.

Note: Wong et al. (2004) and Yang et al. (2005) predictions

LiBaC will be most useful when implemented in conjunction with other approaches. In this way it contributes to a “toolbox” of methods best applied to real data with the goal of studying the process of molecular evolution.

LiBaC appears promising amino acid models, where sites in the folded protein are subject to different microenvironments and perform different functions.
**Amino acid physiochemical properties?**

1. **ATA (Ile) → TTA (Leu):** $\omega_{TTA}$ [conservative]
2. **ATA (Ile) → AAA (Lys):** $\omega_{AAA}$ [radical]

Note: Commonly used models average $\omega$ over physiochemical properties

**Physiochemical properties and estimates of $\omega$**

- Yang et al. 1998
- Sainudiin et al. 2005 [mechanistic]
- Wong et al. 2006
- Kosiol et al. 2007
- Doron-Faigenboim & Pupko 2007 [mechanistic + empirical *]

**NOTES:**
- The $\omega$ is not comparable to mechanistic models without correction.
- Performance of site classification under these models in unknown.
- Might have more immediate use in phylogenetic reconstruction.
1. A likelihood gain + changes in site ID under a “new” model is **not** an adequate indicator of classification performance.

2. Variable $d_s$ among sites is not necessarily a problem for classification of sites. ([also see: Yang & Nielsen. 2008. MBE, 25:568-579.](#))

3. The level of difficulty depends on the data in hand.

4. More careful model selection & model assessment (than is currently being employed) is warranted.

5. **Use M2a**; LRTs and site ID appear to be quite robust.

6. **Risks to real data analysis remain largely unknown!**

---

“We are drowning in information and starving for knowledge.”

—Rutherford D. Roger