codon substitution models and the analysis of natural selection pressure

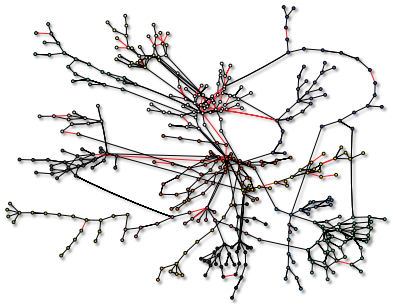
Joseph P. Bielawski
Department of Biology
Department of Mathematics & Statistics
Dalhousie University

codon models have many uses

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

morphological adaptation

metabolic networks



protein structure

Troponin C: fast skeletal

Troponin C: cardiac and slow skeletal

gene sequences

```
human
cow
rabbit
rat
opossum

GTG CTG TCT CCT GCC GAC AAG ACC AAC GTC AAG GCC GCC TGG GGC AAG GTT GGC GCG CAC
... ... ... G.C ... ... ... T.. ..T ... ... ... ... ... ... ... ... .GC A..
... ... ... ..C ..T ... ... ... ... A.. ... A.T ... ... .AA ... A.C ... AGC ...
... ..C ... G.A .AT ... ..A ... ... A.. ... AA. TG. ... ..G ... A.. ..T .GC ..T
... ..C ..G GA. ..T ... ... ..T C.. ..G ..A ... AT. ... ..T ... ..G ..A .GC ...

GCT GGC GAG TAT GGT GCG GAG GCC CTG GAG AGG ATG TTC CTG TCC TTC CCC ACC ACC AAG
... ..A .CT ... ..C ..A ... ..T ... ... ... ... ... AG. ... ... ... ... ... ...
.G. ... ... ... ..C ..C ... ... G.. ... ... ... ... T.. GG. ... ... ... ... ...
.G. ..T ..A ... ..C .A. ... ... ..A C.. ... ... ... ... GCT G.. ... ... ... ... ...
..C ..T .CC ..C .CA ..T ..A ..T ..T .CC ..A .CC ... ..C ... ... ... ..T ... ..A

ACC TAC TTC CCG CAC TTC GAC CTG AGC CAC GGC TCT GCC CAG GTT AAG GGC CAC GGC AAG
... ... ... ..C ... ... ... ... ... ... ... ..G ... ... ..C ... ... ... ... G..
... ... ... ..C ... ... ... T.C .C. ... ... ... .AG ... A.C ..A .C. ... ... ... ...
... ... ... T.T ... A.T ..T G.A ... .C. ... ... ... ... ..C ... .CT ... ... ... ...
..T ... ... ..C ... ... ... ... TC. .C. ... ..C ... ... A.C C.. ..T ..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

overview

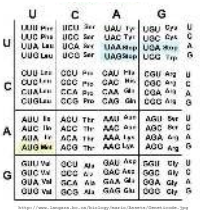
1. The ω ratio (d_N/d_S): the basics
2. Markov models of codon evolution
3. Model based inference
4. Sequence evolution is complex ("know your data")

morning

5. PAML introduction
6. Real data exercises

evening

an index of natural selection pressure



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.

Kimura (1968)

d_S: number of synonymous substitutions per synonymous site (K_S)

d_N: number of nonsynonymous substitutions per nonsynonymous site (K_A)

ω: the ratio d_N/d_S; it measures selection at the protein level

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**)**?**

the basics

Relative proportion of different types of mutations in hypothetical protein coding sequence.

Type	Expected number of changes (proportion)			
	All 3 Positions	1st positions	2nd positions	3rd positions
Total mutations	549 (100)	183 (100)	183 (100)	183 (100)
Synonymous	134 (25)	8 (4)	0 (0)	126 (69)
Nonsyonymous	392 (71)	166 (91)	176 (96)	57 (27)
nonsense	23 (4)	9 (5)	7 (4)	7 (4)

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.

the basics

Why use d_N and d_S?

Same example, but using d_N and d_S:

Synonymous sites = 25.5%
S = 300 × 3 × 25.5% = 229.5

Nonsynonymous sites = 74.5%
N = 300 × 3 × 74.5% = 670.5

So, d_S = 5/229.5 = 0.0218
d_N = 5/670.5 = 0.0075

d_N/d_S (ω) = 0.34, **purifying selection** !!!

mutational opportunity

Relative proportion of different types of mutations in hypothetical protein coding sequence.

	Expected number of changes (proportion)			
Type	**All 3 Positions**	**1st positions**	**2nd positions**	**3rd positions**
Total mutations	549 (100)	183 (100)	183 (100)	183 (100)
Synonymous	134 (25)	8 (4)	0 (0)	126 (69)
Nonsyonymous	392 (71)	166 (91)	176 (96)	57 (27)
nonsense	23 (4)	9 (5)	7 (4)	7 (4)

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**.

real data have biases (*Drosophila GstD1* gene)

transitions vs. **transversions:**

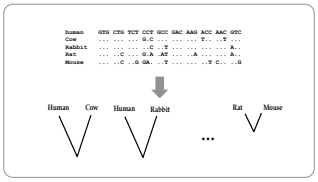
ts/tv = 2.71

preferred vs. **un-preferred** codons:

Partial codon usage table for the *GstD* gene of *Drosophila*

Phe F	TTT	0	Ser S	TCT	0	Tyr Y	TAT	1	Cys C	TGT	0
	TTC	27		TCC	15		TAC	22		TGC	6
Leu L	TTA	0		TCA	0	*** *	TAA	0	*** *	TGA	0
	TTG	1		TCG	1		TAG	0	Trp W	TGG	8
Leu L	CTT	2	Pro P	CCT	1	His H	CAT	0	Arg R	CGT	1
	CTC	2		CCC	15		CAC	4		CGC	7
	CTA	0		CCA	3	Gln Q	CAA	0		CGA	0
	CTG	29		CCG	1		CAG	14		CGG	0

pairwise counting methods



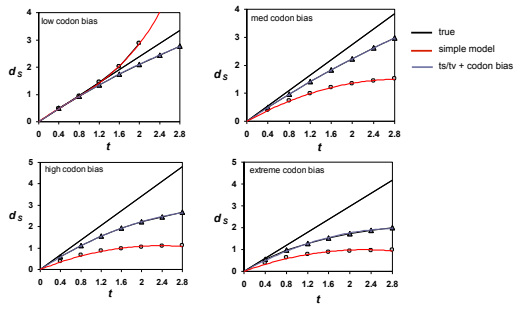
1. Count synonymous (*S*) and nonsynonymous (*N*) sites
2. Count synonymous and nonsynonymous differences
3. Apply some corrections (*e.g.*, correct for multiple hits)

selected counting methods

1. Perler, F. et al. 1980. Cell 20: 555-566
2. Miyata, T. & T. Yasunaga. 1980. *JME* 16:23-36
3. Li, W.-H., C.-I. Wu, & C.-C. Luo. 1985. MBE 2:150-174
4. Nei, M. & T. Gojobori. 1986. *MBE* 3: 418-426
5. Li, W.-H. 1993. *JME* 36:96-99
6. Pamilo & Bianchi 1993 *MBE* 10:271-281
7. Ina, Y. 1995. *JME* 40:190-226
8. Comeron, J. M. 1995. *JME* 41:1152-1159
9. Moriyama, E. N. & F. R. Powell, 1997. *JME* 45:378-391
10. Yang, Z., and R. Nielsen. 2000. *MBE* 17:32-43.

no ts/tv bias + no codon bias
ts/tv bias + no codon bias
ts/tv bias + codon bias

"corrections" and estimation bias in d_S



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

What now?

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!)

model based inference

"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)

Markov models of codon evolution

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

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

"GY-style" codon models

important parameters:

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

"GY-style" codon models

rates to CTG

Synonymous

CT**C** (Leu) → CT**G** (Leu): π_{CTG}

TTC (Leu) → **C**TG (Leu): $\kappa\pi_{CTG}$

Nonsynonymous

GTG (Val) → **C**TG (Leu): $\omega\pi_{CTG}$

C**C**G (Pro) → C**T**G (Leu): $\kappa\omega\pi_{CTG}$

"GY-style" codon models

	to codon below:							
From codon below:	TTT (Phe)	TTC (Phe)	TTA (Leu)	TTG (Leu)	CTT (Leu)	CTC (Leu)	⋯▸	GGG (Gly)
TTT (Phe)	—	$\kappa\pi_{TTC}$	$\omega\pi_{TTA}$	$\omega\pi_{TTG}$	$\omega\kappa\pi_{TTT}$	0	⋯▸	0
TTC (Phe)	$\kappa\pi_{TTT}$	—	$\omega\pi_{TTA}$	$\omega\pi_{TTG}$	0	$\omega\kappa\pi_{CTC}$	⋯▸	0
TTA (Leu)	$\omega\pi_{TTT}$	$\omega\pi_{TTC}$	—		0	0	⋯▸	0
TTG (Leu)	$\omega\pi_{TTT}$	$\omega\pi_{TTC}$	$\kappa\pi_{TTA}$	—	0	0	⋯▸	0
CTT (Leu)	$\omega\kappa\pi_{TTT}$	0	0	0	—	$\kappa\pi_{CTC}$	⋯▸	0
CTC (Leu)	0	$\omega\kappa\pi_{TTC}$	0	0	$\kappa\pi_{TTT}$	—	⋯▸	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	
GGG (Gly)	0	0	0	0	0	0		—

* This is equivalent to the codon model of Goldman and Yang (1994). Parameter ω is the ratio d_N/d_S. κ is the transition/transversion rate ratio, and π_i is the equilibrium frequency of the target codon (*i*).

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

"GY-style" codon models

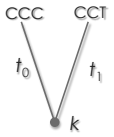
(Goldman & Yang 1994 *MBE* **11**:725-736)

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

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

codon models

likelihood of the data at a site (only two codons)



$$L_h(CCC, CCT) = \sum_k \pi_k p_{kCCC}(t_0) p_{kCCT}(t_1)$$

Note: analysis is typically done by using an unrooted tree

codon models

likelihood of the data at all sites

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.

$$\ell = \ln\{L\} = \ln\{L_1\} + \ln\{L_2\} + \ln\{L_3\} + \ldots + \ln\{L_N\} = \sum_{h=1}^{N} \ln\{L_h\}$$

codon models

we made some progress …

the good: we now have a framework for …

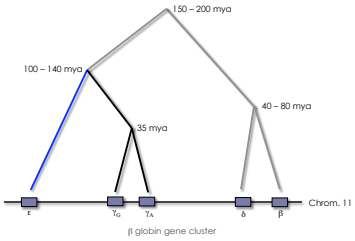
- avoiding *ad hoc* corrections of counting methods
- computation of transition probabilities *
- principled framework for statistical inference

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**" $\omega > 1$!

* 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.

codon models

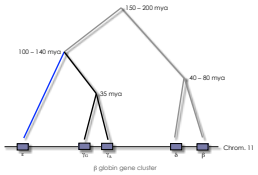
selection pressure (ω) varies in time



Our question: **When?**

codon models

selection pressure (ω) varies in time



b.l. (my)	Fraction of *t*	ω
115	0.203	0.2
55	0.097	0.2
60	0.106	0.2
60	0.106	0.2
85	**0.150**	**0.5**
35	**0.062**	**0.5**
35	**0.062**	**0.5**
120	**0.212**	**1.2**
(Tot = 565my)	(Tot = 1)	

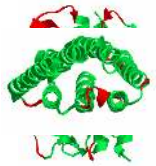
Grey branches: ω = 0.2
Black branches: ω = 0.5
Blue branches: ω = 1.2

If we average over the tree, we do NOT detect positive selection;

$\omega = 0.49$.

codon models

selection pressure (ω) varies among sites



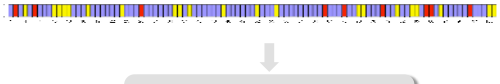
Our question: **Where?**

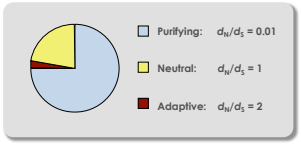
codon models

selection pressure (ω) varies among sites

If we average over sites, we do NOT detect positive selection;

$\omega = 0.31$



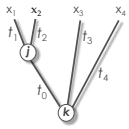


codon models

likelihood of a phylogeny

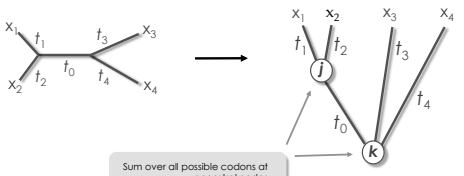
problem: averaging ω over a pair has very low power if the questions are about "**when**" or "**where**"!

solution: phylogenetic estimation of selection pressure



codon models

likelihood of a phylogeny



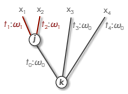
Sum over all possible codons at ancestral nodes

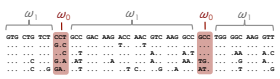
{TTT, TTC, …, GGG}

$$L(x_h) = \sum_k \left[\pi_k p_{kx_4}(t_4) p_{kx_3}(t_3)\right.$$

codon models

we made some more progress ….





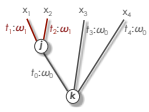
1. branch models
(ω varies among branches)

2. site models
(ω varies among sites)

3. branch-site models
(combines the features of above models)

codon models

1. branch models *

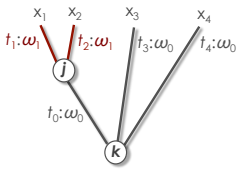


Variation (ω) among branches:	Approach
Yang, 1998	fixed effects
Bielawski and Yang, 2003	fixed effects
Seo et al. 2004	auto-correlated rates
Kosakovsky Pond and Frost, 2005	genetic algorithm
Dutheil et al. 2012	clustering algorithm

* *These methods can be useful when selection pressure is strongly episodic*

codon models

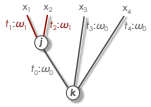
1. branch models



$$L(x_h) = \sum_k \sum_j \left[\pi_k p_{kx_4}(t_4;\omega_0) p_{kx_3}(t_3;\omega_0) p_{kj}(t_0;\omega_0) p_{jx_2}(t_2;\omega_1) p_{jx_1}(t_1;\omega_1)\right]$$

codon models

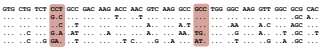
1. branch models



- 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

codon models

2. site models *

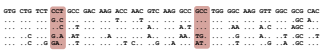


Variation (ω) among sites:	Approach
Yang and Swanson, 2002	fixed effects (ML)
Bao, Gu and Bielawski, 2006	fixed effects (ML)
Massingham and Goldman, 2005	site wise (LRT)
Kosakovsky Pond and Frost, 2005	site wise (LRT)
Nielsen and Yang, 1998	mixture model (ML)
Kosakovsky Pond, Frost and Muse, 2005	mixture model (ML)
Huelsenbeck and Dyer, 2004	mixture (Bayesian)
Huelsenbeck et al. 2006	mixture (Bayesian)
Bao, Gu, Dunn and Bielawski 2008	mixture (LiBaC/MBC)

- *Useful when at some sites evolve under diversifying selection pressure over long periods of time*
- *This is not a comprehensive list.*

codon models

2. site models



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

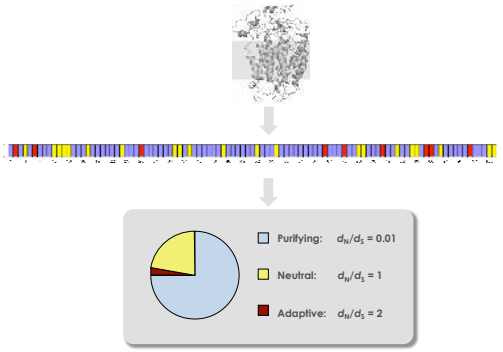
codon models

2. site models: "M-series"

Model	Code	NP	Parameters
One-ratio	M0	1	ω
Neutral	M1a	2	$p_0, \omega_0,$
Selection	M2a	4	$p_0, p_1, \omega_0, \omega_2$
Discrete	M3	2K-1	$p_0, p_1, ..., p_{K-2}$ $\omega_0, \omega_1, ..., \omega_{K-2}$
Frequency	M4	5	$p_0, p_1, ... p_4$
Gamma	M5	2	α, β
2Gamma	M6	4	$p_0, \alpha_0, \beta_0, \alpha_1$
Beta	M7	2	p, q
Beta&ω	M8	4	p_0, p, q, ω
Beta&gamma	M9	5	p_0, p, q, α, β
Beta&normal+1	M10	5	p_0, p, q, α, β
Beta&normal>1	M11	5	p_0, p, q, μ, σ
0&2normal>1	M12	5	$p_0, p_1, \mu_2, \sigma_1, \sigma_2$
3normal>0	M13	6	$p_0, p_1, \mu_2, \omega_3, \sigma_1, \sigma_2$

codon models

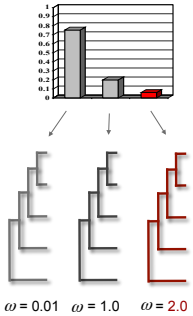
2. site models: ω varies among sites



codon models

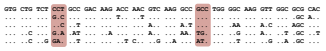
2. site models: M3 (discrete)

$$P(\mathbf{x}_h) = \sum_{i=0}^{K-1} p_i P(\mathbf{x}_h \mid \omega_i)$$



codon models

2. 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

codon models

3. branch-site models ++

Variation (ω) among branches & sites:	Approach
Yang and Nielsen, 2002	fixed (ML)
Forsberg and Christiansen, 2003	fixed (ML)
Bielawski and Yang, 2004	fixed (ML)
Giundon et al., 2004	switching (Bayesian)
Zhang et al. 2005	mixture (ML)
Kosakovsky Pond et al. 2011, 2012	mixture (ML)

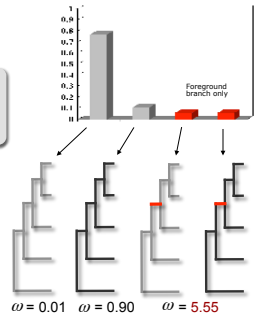
* *These methods can be useful when selection pressures change over time at just a fraction of sites*

* *It can be a challenge to apply these methods properly.*

codon models

3. branch-site model: "model B"

$$P(\mathbf{x}_h) = \sum_{i=0}^{K-1} p_i P(\mathbf{x}_h \mid \omega_i)$$



ω for background branches are from site-classes 1 and 2 (0.01 or 0.90)

codon models

lots of options …

1. **Pairwise methods** have very low power to detect adaptive evolution (via counting or likelihood method), but they are very fast.
2. **Branch models** allow variation among branches but assume one ω for all sites, and have low power to detect positive selection. Good for detecting shifts in average selection intensity.
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!)

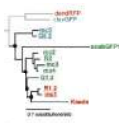
10:30?

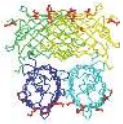
Let's take a break

model based inference

colour diversity of coral pigments

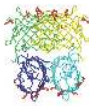
Red/blue colour morphs of the great star coal *Montastraea cavernosa*





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

green fluorescent proteins (GFPs) in corals



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?

codon models

analytical tasks

Task 1. Parameter estimation (e.g., ω) ⬅

Task 2. Hypothesis testing

Task 3. Make predictions (e.g., sites having $\omega > 1$)

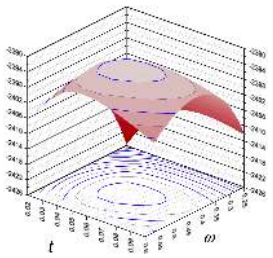
ML parameter estimation

t, κ, ω = unknown values estimated by ML

π's = empirical [GY: F3×4 or F61 in Lab]

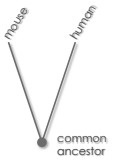
use a numerical hill-climbing algorithm to maximize the likelihood function

ML parameter estimation



Parameters: t and ω

Gene: acetylcholine α receptor



lnL = -2399

How do we know that the estimate is significant?

Task 1. Parameter estimation (e.g., ω) ✔

Task 2. Hypothesis testing ⬅ **LRT**

Task 3. Prediction / Site identification

testing nested hypotheses by using the LRT

ℓ_0 is the maximum log likelihood under $\mathbf{H_0}$ with parameters $\boldsymbol{\theta_0}$

and

ℓ_1 is the maximum log likelihood under $\mathbf{H_1}$ with parameters $\boldsymbol{\theta_1}$

$$\text{Test statistic} = 2\Delta\ell = 2(\ell_0(\theta_0) - \ell_1(\theta_1))$$

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

Task 2

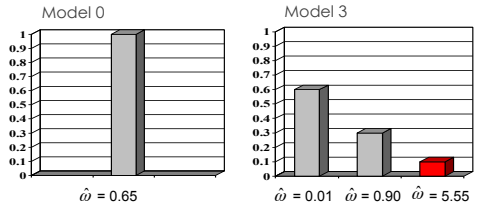
testing nested hypotheses by using the LRT

Model	Code	NP	Parameters
One-ratio	M0	1	ω
Neutral	M1a	2	p_0, ω_0
Selection	M2a	4	$p_0, p_1, \omega_0, \omega_2$
Discrete	M3	2K-1	$p_0, p_1, \ldots p_{K-2}$ $\omega_0, \omega_1, \ldots, \omega_{K-2}$
Frequency	M4	5	$p_0, p_1, \ldots p_4$
Gamma	M5	2	α, β
2Gamma	M6	4	$p_0, \alpha_0, \beta_0, \alpha_1$
Beta	M7	2	p, q
Beta&ω	M8	4	p_0, p, q, ω
Beta&gamma	M9	5	p_0, p, q, α, β
Beta&normal+1	M10	5	p_0, p, q, α, β
Beta&normal>1	M11	5	p_0, p, q, μ, σ
0&2normal>1	M12	5	$p_0, p_1, \mu_2, \sigma_1, \sigma_2$
3normal>0	M13	6	$p_0, p_1, \mu_2, \sigma_0, \sigma_1, \sigma_2$

Task 2

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

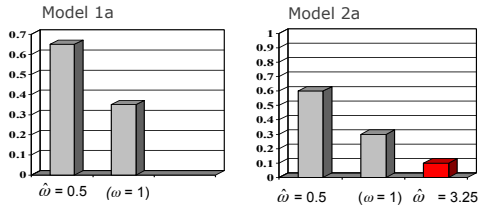
H_0: uniform selective pressure among sites (M0)
H_1: variable selective pressure among sites (M3)
Compare **$2\Delta l$** = $2(l_1 - l_0)$ with a χ^2 distribution



Task 2

LRT No. 2: Have some sites evolved under positive selection?

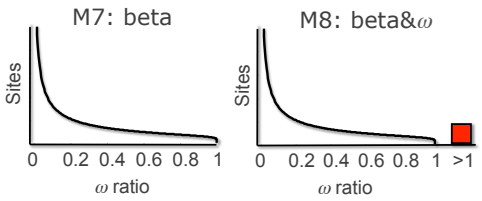
H_0: variable selective pressure but NO positive selection (M1)
H_1: variable selective pressure with positive selection (M2)
Compare **$2\Delta l$** = $2(l_1 - l_0)$ with a χ^2 distribution



Task 2

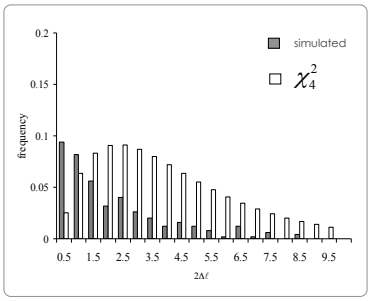
LRT No. 3: Have some sites evolved under positive selection?

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 χ^2 distribution



Task 2

the LRT does not follow the χ^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 α = 1% (5%) significance levels

Exp.	Simulation	LRT	Simulation parameters				Type I error at α = 1% (5%)	
			Taxa	κ	ω	t	N = 100	N = 500
A....	M0	M0 & M3	6	2	0.40	0.11	0 (0)	0 (0)
						1.1	0 (0)	0 (0)
						11	0 (0)	0 (0)
B....	M0	M0 & M3	17	2	0.40	2.11	0 (0)	0 (1)
						8.44	0 (0)	0 (1)
						16.88	0 (1)	0 (0)
C...	M0	M0 & M3	5	5	0.25	0.91	0 (0)	0 (0)
						9.1	0 (0)	0 (1)
						18.2	0 (1)	2 (3)
D...	M7	M7 & M8	6	2	p = 0.41 q = 1.10	0.11	N/A	0 (0)
						1.1	N/A	1 (5)
						11	N/A	1 (4)

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

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

Task 2

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

Simulation	LRT	Simulation parameters				P_+		$P_{+,\,0.01\,(0.05)}$	
		Taxa	κ	ω distribution	t	L_C = 100	L_C = 500	L_C = 100	L_C = 500
M3	M0 & M3	17	2	ω_0 = 0.018, p_0 = 0.386 ω_1 = 0.304, p_1 = 0.535 **ω_2 = 1.691, p_2 = 0.079**	**0.38**	61	80	10 (17)	66 (72)
					2.11	93	100	91 (92)	100 (100)
					8.44	99	100	99 (99)	100 (100)
					16.88	99	99	99 (99)	99 (99)
					105.5	31	58	31 (31)	58 (58)

NOTE: Here t 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., ω) ✔

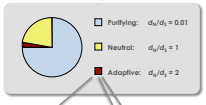
Task 2. Hypothesis testing ✔

Task 3. Prediction / Site identification ← Bayes' rule

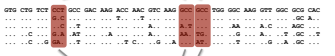
Task 3

Which sites have $\omega > 1$?

model:
5% have $\omega > 1$



Bayes' rule:
site 4, 12 & 13



structure:
sites are in contact

Task 3

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

$$P(\text{D}) = P(\text{M})P(\text{D}|\text{M}) + P(\text{F})P(\text{D}|\text{F})$$

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

Task 3

Bayes' rule: yet another (silly) example of

Q_2: Given that an individual carries the disease, what is the probability that it is a male?

A_2: $0.6 \times 0.01/0.0064 = 0.94$

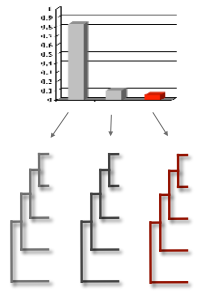
$$P(\text{M}|\text{D}) = \frac{P(\text{M})\,P(\text{D}|\text{M})}{P(\text{D})}$$

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

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



ω_0 = 0.03 ω_1 = 0.40 ω_2 = 14.1
p_0 = 0.85 p_1 = 0.10 p_2 = 0.05

Task 3

Bayes' rule for identifying selected sites

Site class 0: ω_0 = .03, 85% of codon sites

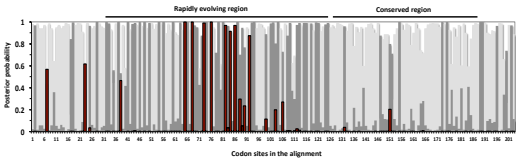
Site class 1: ω_1 = .40, 10% of codon sites

? → **Site class 2: ω_2 = 14, 05% of codon sites**

$$P(\omega_2 \mid \mathbf{x}_h) = \frac{p_2 P(x_h \mid \omega_2)}{P(x_h)} = \frac{p_2 P(x_h \mid \omega_2)}{\sum_{i=0}^{K-1} p_i P(x_h \mid \omega_i)}$$

Task 3

Bayes' rule for identifying selected sites



Site class 0: ω_0 = .03 (strong purifying selection)

Site class 1: ω_1 = .40 (weak purifying selection)

Site class 2: ω_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

Task 3

Bayes' rule for identifying selected sites

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 ω distribution parameters
2. Fix branch lengths to MLEs
3. Integrate over uncertainty
4. BEB is faster than "Full Bayes" (FB)

	False classification rates		
Small datasets:	FB/BEB	<	NEB
Large datasets:	FB/BEB	≈	NEB*

* exception: extreme parameter estimates

See: Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22(4):1107-1118.

model based inference

even more progress …

Task 1. Parameter estimation (e.g., ω) ✔

Task 2. Hypothesis testing ✔

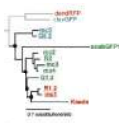
Task 3. Prediction / Site identification ✔

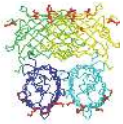
let's put this into practice …

model based inference

colour diversity of coral pigments

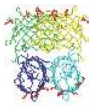
Red/blue colour morphs of the great star coal *Montastraea cavernosa*





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

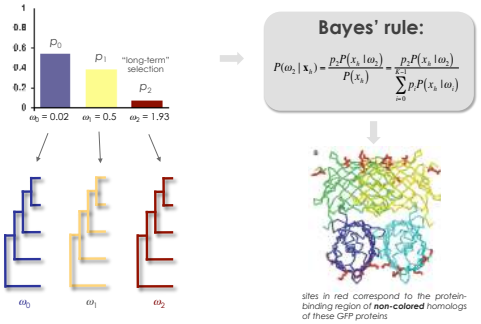
green fluorescent proteins (GFPs) in corals



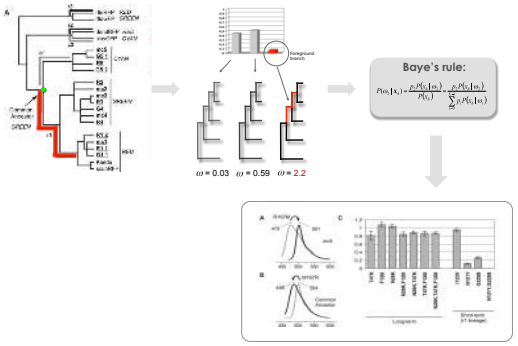
Questions we have:

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?

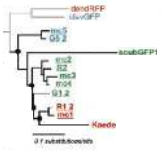
signal 1: long term (diversifying) selection

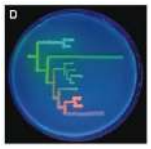


signal 2: episodic selection



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

model based inference

"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 !!!

"know your data"

sequence evolution is complex

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

"know your data"

how to model codon frequencies?

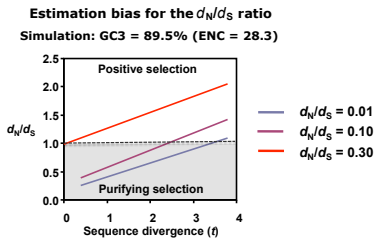
Sums of codon usage counts the *GstD* gene of *Drosophila*

Phe	F	TTT	0	Ser	S	TCT	0	Tyr	Y	TAT	1	Cys	C	TGT	0
		TTC	27			TCC	15			TAC	22			TGC	6
Leu	L	TTA	0			TCA	0	***	*	TAA	0	***	*	TGA	0
		TTG	1			TCG	1			TAG	0	Trp	W	TGG	8
Leu	L	CTT	2	Pro	P	CCT	1	His	H	CAT	0	Arg	R	CGT	1
		CTC	2			CCC	15			CAC	4			CGC	7
		CTA	0			CCA	3	Gln	Q	CAA	0			CGA	0
		CTG	29			CCG	1			CAG	14			CGG	0
Ile	I	ATT	4	Thr	T	ACT	2	Asn	N	AAT	5	Ser	S	AGT	1
		ATC	12			ACC	11			AAC	17			AGC	4
		ATA	0			ACA	2	Lys	K	AAA	1	Arg	R	AGA	0
Met	M	ATG	4			ACG	4			AAG	37			AGG	1
Val	V	GTT	0	Ala	A	GCT	0	Asp	D	GAT	2	Gly	G	GGT	4
		GTC	2			GCC	38			GAC	11			GGC	6
		GTA	1			GCA	2	Glu	E	GAA	0			GGA	11
		GTG	25			GCG	3			GAG	30			GGG	0

"know your data"

biological conclusions are affected

positive selection in highly biased gene sequences?

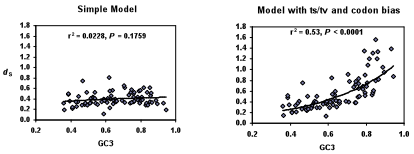


Unpublished simulation study

"know your data"

biological conclusions are affected

what is the genomic relationship between d_S and GC content in mammals ?



Artiodactyla vs. Primates (82 nuclear genes)

Data from: Bielawski, Dunn, and Yang (2000) *Genetics*: 156: 1299-130)

"know your data"

how to model codon frequencies?

	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.

"know your data"

how to model codon frequencies?

Example: A → C

AAA → CAA

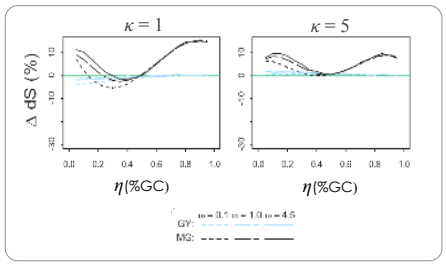
AAA → ACA

AAA → AAC

	Δ at codon position		
	1st	2nd	3rd
GY	π_{CAA}	π_{ACA}	π_{AAC}
MG	$\pi_c{}^1$	$\pi_c{}^2$	$\pi_c{}^3$

biological conclusions are affected

Can choice of frequency model impact biological conclusions?



Data from: Aris-Brosou and Bielawski (2006) Gene, 378:58-64

conditional nucleotide frequencies (CNF)

Example: A → C

AAA → CAA

AAA → ACA

AAA → AAC

	Δ at codon position		
	1st	2nd	3rd
GY	π_{CAA}	π_{ACA}	π_{AAC}
MG	π_c^1	π_c^2	π_c^3
CNF	$\pi_c^1 \mid \pi_{AA}^{2,3}$	$\pi_C^2 \mid \pi_{A,A}^{1,3}$	$\pi_c^3 \mid \pi_{AA}^{1,2}$

For more details: Yap et al. (2010) mbe 27: 726-734.

how to model codon frequencies?

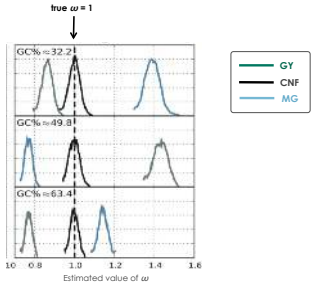


Figure adapted from Yap et al. (2010) mbe 27: 726-734.

how to model codon frequencies?

Caution: Inadequate, or inappropriate, modeling of codon usage can lead to estimation errors for the ω 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**

sequence evolution is complex

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

sequence evolution is complex

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

Should we be concerned?

"know your data"

transmenbrane proteins

Partition sites within a gene:

Loop structures extend into extra-cellular space: **Hydrophilic amino acids here** — $\omega_0\ \pi_0\ \kappa_0\ c_0$

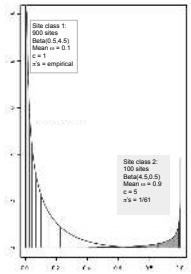
Cell membrane in grey; helix structures span the membrane: **Hydrophobic amino acids here** — $\omega_1\ \pi_1\ \kappa_1\ c_1$

Loop structures extend into cytoplasm: **Hydrophilic amino acids here** — $\omega_2\ \pi_2\ \kappa_2\ c_2$

GY-type codon models: variable ω' + c's among sites = variable d_N & d_S among sites

"know your data"

Realistic simulation w/ no positive selection



	error	site class 1	site class 2
M2a*†	7.2% (7.7%)	ω_0 = 0.06; [w_1 = 1] p_0 = 0.83; p_1 = 0.13	**ω_2 = 2.48** [p_2 = 0.04]
M3†	6.4%	ω_1 = 0.07 p_1 = 0.86	**ω_2 = 1.45** [p_2 = 0.14]
M8*†	5.5% (6.0%)	p = 0.33, q = 2.51 p_1 = 0.91	**ω_2 = 1.92** [p_2 = 0.09]
Soft LiBaC 1	10.1%	ω_1 = 0.04 {p_1 = 0.79}	ω_2 = 0.72 {p_2 = 0.21}
Soft LiBaC 2	9.6%	ω_1 = 0.05 {p_1 = 0.80}	ω_2 = 0.72 {p_2 = 0.20}

* LRTs for positive selection were highly significant in all cases.

† High posterior probabilities for sites having ω > 1 in all replications.

"know your data"

biological conclusions are affected

A sample of eight transmembrane proteins from *Rickettsia*. Results are for the group of sites having the largest value of ω.

		Inference of positive selection (*sig. LRT)			
	Nc	**M2a**	**M8**	**LiBaC-1**	**LiBaC-2**
Trbl-VirB6_3	938	***ω = 5.8****	***ω = 4.3****	ω = 0.43	ω = 0.42
Rfal	403	***ω = 4.29****	***ω = 3.35****	***ω = 1.73***	***ω = 1.31***
Ccmf	635	***ω = 15.5****	***ω = 5.57****	***ω = 1.07***	***ω = 2.80***
nuoL3	499	***ω = 12.5****	***ω = 10.4****	***ω = 1.45***	***ω = 1.39***
Trbl-VirB6_2	657	***ω = 32.8***	***ω = 1.79****	ω = 0.44	ω = 0.45
perM	351	***ω = 2.57***	***ω = 2.91 ****	ω = 0.26	ω = 0.10
mivN	504	***ω = 5.95***	***ω = 2.52****	ω = 0.15	ω = 0.18
pgpA	198	***ω = 35.0***	***ω = 3.60****	ω = 0.57	ω = 0.31

"know your data"

sequence evolution is complex

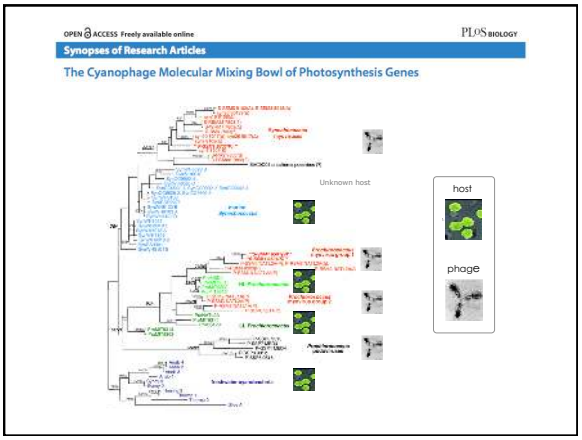
Caution: Inadequate modeling of the complexity of among sites variation can lead to estimation biases.

- un-modeled site-variability can be "soaked-up" by other parameters free to vary among sites
- if ω 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**

"know your data"

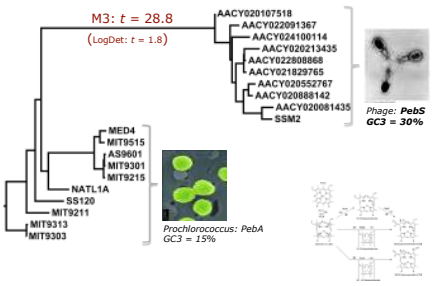
sequence evolution is complex

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



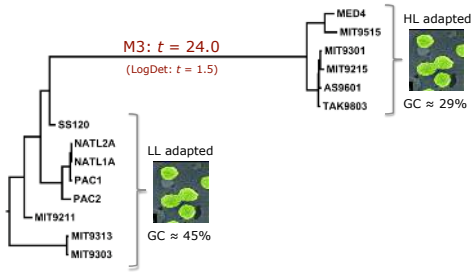
"know your data"

pebS: evolution of a new phage gene



"know your data"

HL and LL adapted *Prochlorococcus* ecotypes



"know your data"

@!#%^! … that's a lot of substitutions

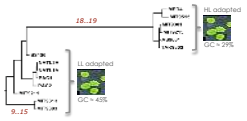


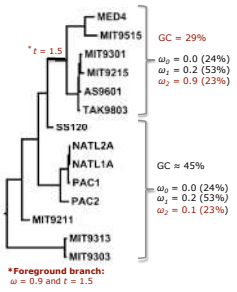
Table: Highly biased branch-specific summary statistics under M0

Branch		t	N	S	d_N	d_S	$N*d_N$	$S*d_S$
18..19	Fequal	6.118	375.3	134.7	0.5244	6.2595	196.8	843.4
	F3x4	10.461	401.2	108.8	0.5446	14.3384	218.5	1559.9
	F61	10.416	411.7	98.3	0.5504	15.7089	226.6	1544.2
9..15*	Fequal	1.622	400.8	139.2	0.0648	1.9111	26	266
	F3x4	8.07	409.1	130.9	0.0624	10.9047	25.5	1427.1
	F61	15.343	413.2	126.8	0.0611	21.5828	25.3	2736.4

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

"know your data"

simulate under the null (no positive selection)



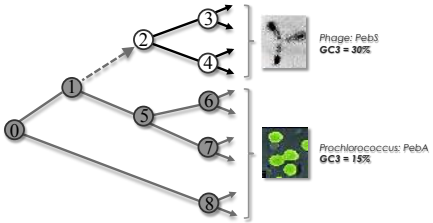
LTR results:

Null	Alternate	% P<0.05
M1	Model A	100
Model A(ω=1)	Model A	30 (60)
M3(k=2)	Model B	100

MLE Results: 100% of estimated values of ω were >1 in the foreground branch

"know your data"

the ulgy: non-stationary models



lower conditional likelihood (lcL)

rooted conditional likelihood (rcL)

$$L_{tree} = \sum_{x_0}\left\{rcL_0 \times \sum_{x_1}\left[Px_0x_1(t_{0,1}) \times \sum_{x_2}\left[Px_1x_2(t_{1,2}) \times lcL_2(x_2)\right] \times \sum_{x_3}\left\{Px_1x_3(t_{1,3}) \sum_{x_5}\left[Px_3x_5(t_{3,5}) \times lcL_5\right] \times \sum_{x_4}\left[Px_3x_4(t_{3,4}) \times lcL_4\right]\right\}\right]\right\}$$

"know your data"

sequence evolution is complex

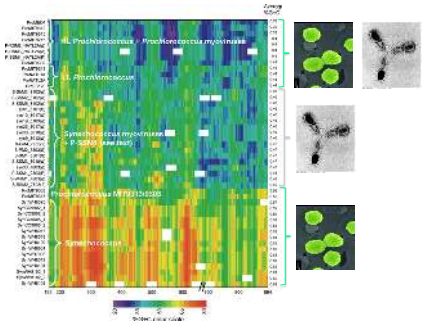
Caution: Serious estimation errors can arise when the process of evolution does not follow the same stochastic pattern in different lineages.

- 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

sequence evolution is complex

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

recombination



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

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:
Anisimova, Nielsen and Yang (2003) Genetics, 164:1229-1236.

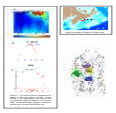
For a solution:
Scheffler et al. (2006) Bioinformatics, 22:2493-2499.

"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"

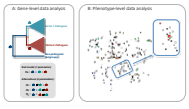
—Andreas Buja

- carry out biophysical analysis of amino acid substitutions at selected sites ("the gold standard")

- test for relationship between adaptive genetic variation and environmental variables



- test for relationship between adaptive genetic variation and complex organism phenotypes



"We are drowning in information and starving for knowledge"

— Rutherford D. Roger