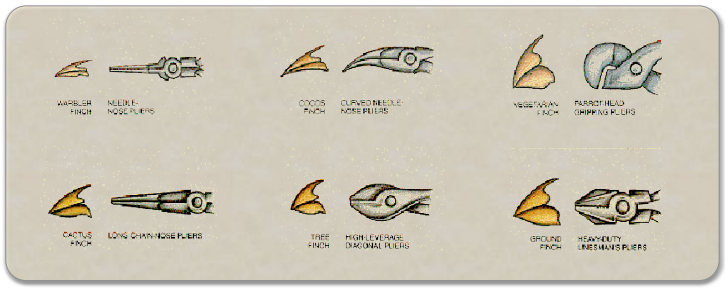
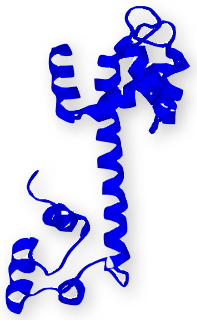
codon substitution models and the analysis of natural selection pressure

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

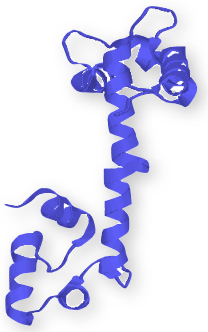
morphological adaptation



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 ...
```

Powerful analytical tools:

1. Population genetic data
2. Comparative analysis of codon sequences
3. Comparative analysis of amino acid sequences

> *"there is no single statistic which is best for testing the most general departures from neutrality"*
>
> (Watterson 1977)

overview

1. introduction to modeling codon evolution
2. model based inference
3. PAML introduction & real data exercises

part I

outline

1. introduction to the ω ratio
2. markov model of codon evolution
3. codon models for ω variation over branches & sites
4. model realism vs. model complexity

1. the ω ratio

an index of natural selection pressure

	U	C	A	G	
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

http://www.langara.bc.ca/biology/mario/Assets/Geneticode.jpg

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

index of natural selection pressure: ω ratio

rate ratio	mode	example
$\omega < 1$	purifying (negative) selection	histones
$\omega = 1$	Neutral Evolution	pseudogenes
$\omega > 1$	Diversifying (positive) selection	MHC, Lysin

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.

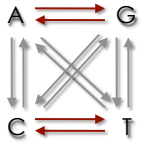
Type	Expected number of changes (proportion)			
	All 3 Positions	1st positions	2nd positions	3rd positions
Total mutations	549 (100)	183 (100)	183 (100)	183 (100)
Synonymous	134 (25)	8 (4)	0 (0)	126 (69)
Nonsyonymous	392 (71)	166 (91)	176 (96)	57 (27)
nonsense	23 (4)	9 (5)	7 (4)	7 (4)

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.**

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**:



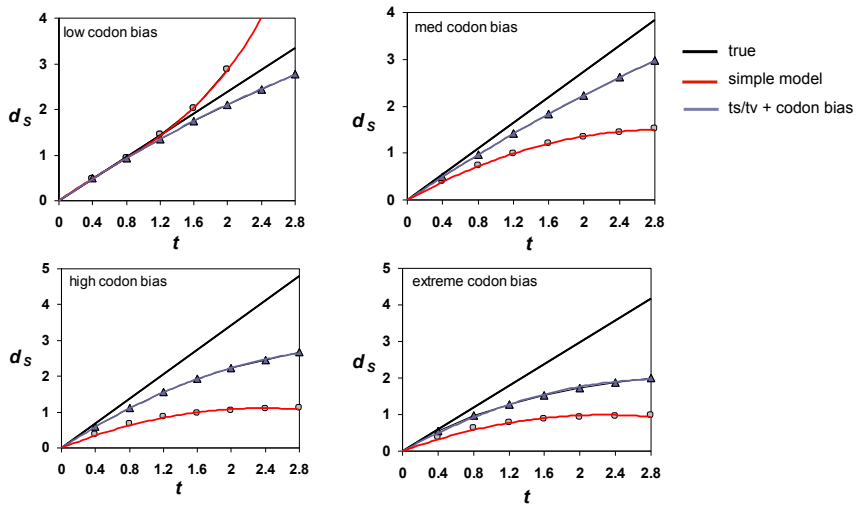
ts/tv = 2.71

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

"corrections" and estimation bias in d_S



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

Markov models of codon evolution

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

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

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

"GY-style" codon models (mechanistic)

some 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 (mechanistic)

let's model a change 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 (mechanistic)

From codon below:	to 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	0	—

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

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

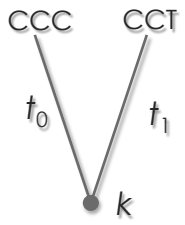
"GY-style" codon models (mechanistic)

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

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

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

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

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.

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

we made some progress …

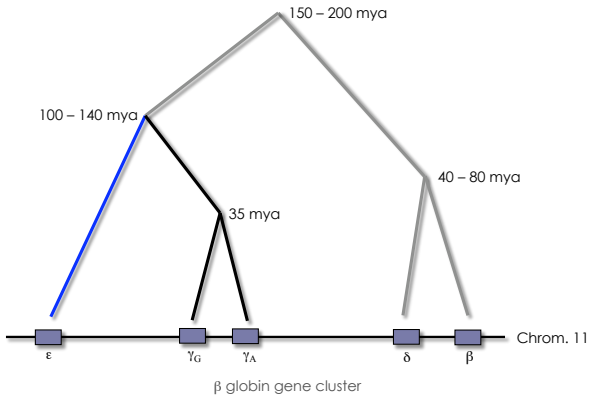
the good: we now have a framework for …

- 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.

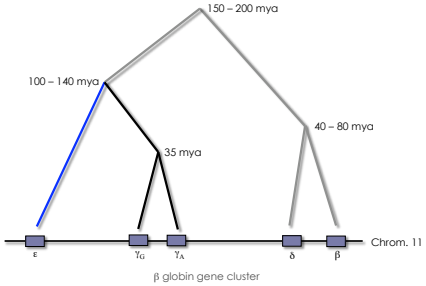
selection pressure (ω) varies in time



Our question: **When?**

2. markovian 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

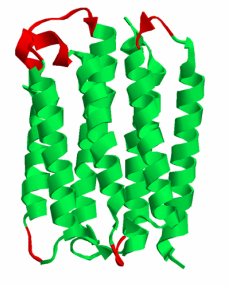
Blue branches: ω = 1.2

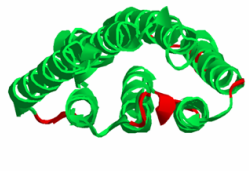
If we average over the tree, we do NOT detect positive selection;

ω = 0.49.

2. markovian codon models

selection pressure (ω) varies among sites



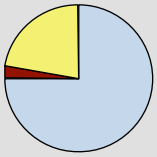


Our question: **Where?**

selection pressure (ω) varies among sites

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

$\omega = 0.31$



Purifying: $d_N/d_S = 0.01$

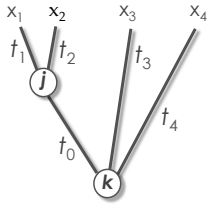
Neutral: $d_N/d_S = 1$

Adaptive: $d_N/d_S = 2$

likelihood of a phylogeny

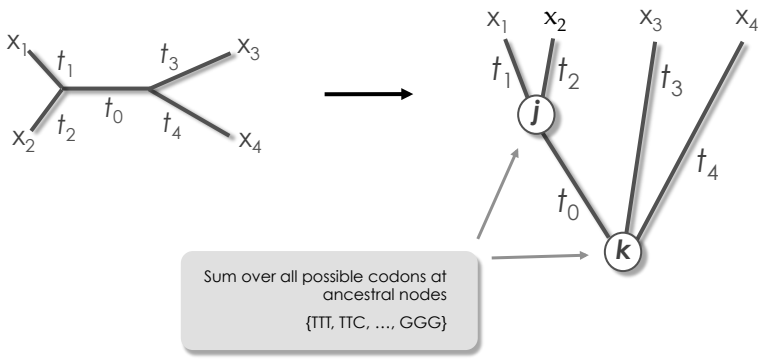
problem: averaging ω over a pair has very low power if the questions are about "**when**" or "**where**"!

solution: phylogenetic estimation of selection pressure



2. markovian codon models

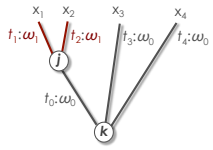
likelihood of a phylogeny



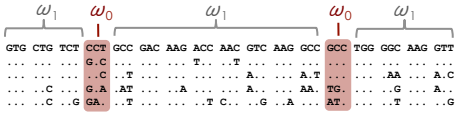
$$L(x_h) = \sum_k \sum_j \left[\pi_k p_{kx_4}(t_4) p_{kx_3}(t_3) p_{kj}(t_0) p_{jx_2}(t_2) p_{jx_1}(t_1) \right]$$

3. ω variation

improvements ….



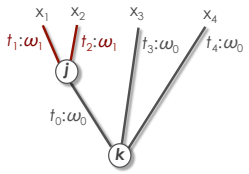
3.1. branch models
(ω varies among branches)



3.2. site models
(ω varies among sites)

3. ω variation

3.1. branch models *

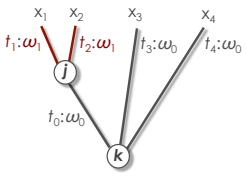


Variation (ω) among branches:	Approach
Yang, 1998	fixed effects
Bielawski and Yang, 2003	fixed effects
Seo et al. 2004	auto-correlated rates
Kosakovsky Pond and Frost, 2005	genetic algorithm
Dutheil et al. 2012	clustering algorithm

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

3. ω variation

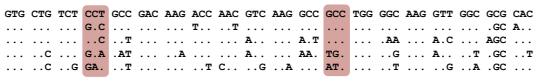
3.1. branch models



- 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

3. ω variation

3.2. 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; Huelsenbeck et al. 2006	mixture (Bayesian)
Rubenstein et al. 2011	mixture model (ML)
Bao, Gu, Dunn and Bielawski 2008 & 2011	mixture (LiBaC/MBC)
Murell et al. 2013	mixture (Bayesian)

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

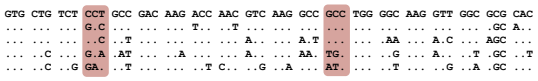
3. ω variation

3.2. site models: "M-series"

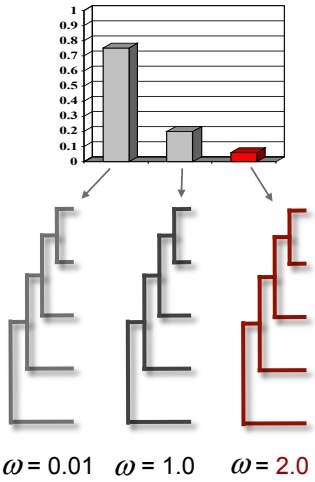
Model	Code	NP	Parameters
One-ratio	M0	1	ω
Neutral	M1a	2	p_0, ω_0
Selection	M2a	4	$p_0, p_1, \omega_0, \omega_2$
Discrete	M3	2K-1	$p_0, p_1, \ldots, p_{K-2}$ $\omega_0, \omega_1, \ldots, \omega_{K-2}$
Frequency	M4	5	$p_0, p_1, \ldots, p_4$
Gamma	M5	2	α, β
2Gamma	M6	4	$p_0, \alpha_0, \beta_0, \alpha_1$
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, α
0&2normal>1	M12	5	$p_0, p_1, \mu_2, \alpha_1, \alpha_2$
3normal>0	M13	6	$p_0, p_1, \mu_2, \alpha_0, \alpha_1, \alpha_2$

3. ω variation

3.2. discrete model



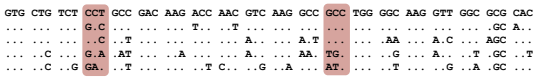
$$P(\mathbf{x}_h) = \sum_{i=0}^{K-1} p_i P(\mathbf{x}_h \mid \omega_i)$$



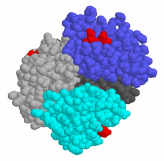
$\omega = 0.01$ $\omega = 1.0$ $\omega = 2.0$

3. ω variation

3.2. site models

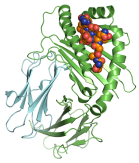


"globin model"



directional selection

"MHC model"



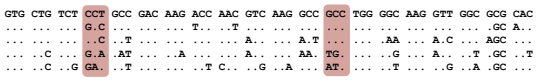
diversifying selection

site models

Site models detect diversifying positive selection; **absence of "signal" under a sites model does not mean changes in a protein did not have fitness consequences** (Bielawski et al. 2004 PNAS 101: 14824–29)

3. ω variation

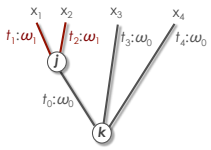
3.2. site models

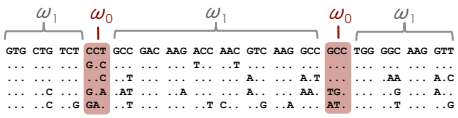


- genetic incompatibilities in human infertility
- non-hormonal contraception drugs
- identify pathogenicity genes
- venom-anti-venom co-evolution
- identify candidate genes for drug therapies
- identify immune and defense system genes
- vaccine design
- aid functional classification of unknown genes
- incorporate in models of protein 2D and 3D structure

3. ω variation

we made some more progress ….





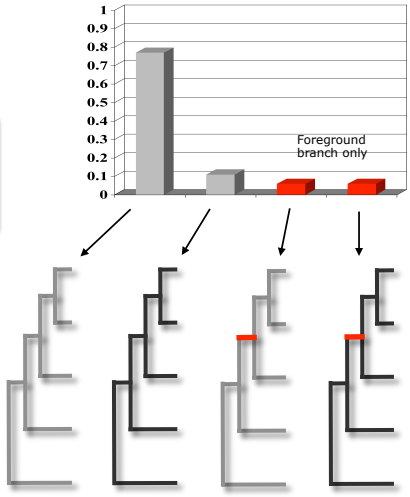
3.1. branch models
(ω varies among branches)

3.2. site models
(ω varies among sites)

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

3.3. branch-site model: "model B"

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



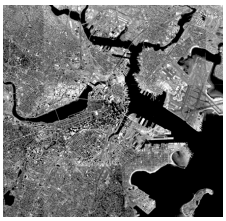
3.3. models for variation in branches & sites

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

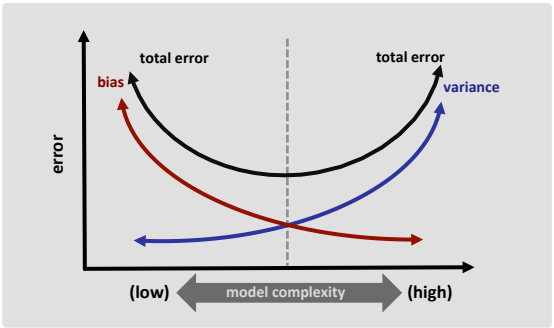
> "a **MODEL** is an intentional simplification of a complex situation designed to eliminate extraneous detail in order to focus attention on the essentials of the situation"
>
> (Daniel L. Hartl)



(images from Paul Lewis)

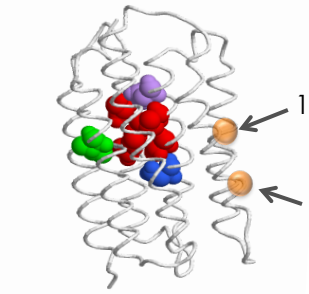
there is a kind of "*tension*" between model realism and model complexity

"UNDER FIT" "OVER FIT"



AA exchangeabilties

intentional simplification: all of the models listed above treat these two amino acid substitutions in the same way!



1. **A**TA (**Ile**) → **T**TA (**Leu**): $\omega\pi_{TTA}$ (**conservative**)

2. ATA (**Ile**) → A**A**A (**Lys**): $\omega\pi_{AAA}$ (**radical**)

Q matrix for M0 codon models

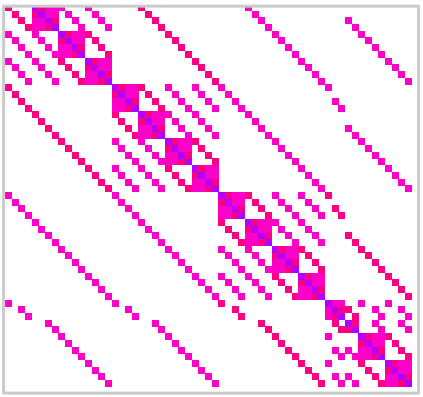
	to codon below:							
From codon below:	**TTT (Phe)**	**TTC (Phe)**	**TTA (Leu)**	**TTG (Leu)**	**CTT (Leu)**	**CTC (Leu)**	⋯▸	**GGG (Gly)**
TTT (Phe)	——	$\kappa\pi_{TTC}$	$\omega\pi_{TTA}$	$\omega\pi_{TTG}$	$\omega\kappa\pi_{TTT}$	0	⋯▸	0
TTC (Phe)	$\kappa\pi_{TTT}$	——	$\omega\pi_{TTA}$	$\omega\pi_{TTG}$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

4. realism vs. complexity

structure of rate matrix: GY-M0

2 parameters (κ and ω)



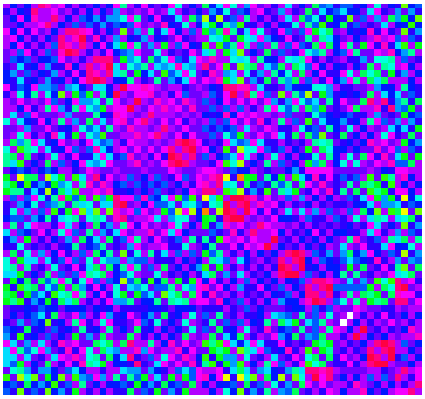
Structure of a Q matrix (log scaled and binned) derived from M0 for Abolone sperm lysin

too sparse?

4. realism vs. complexity

structure of rate matrix: ECM

1830 parameters

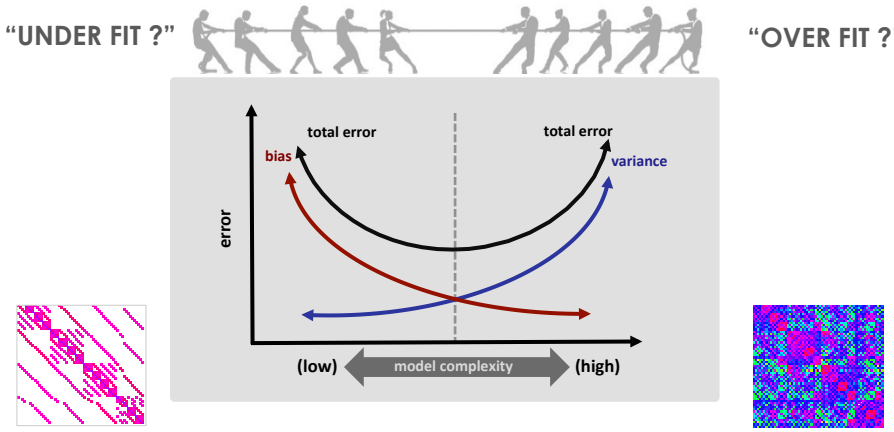


Structure of a Q matrix (log scaled and binned) derived ECM (Kosiol et al. 2007) for all sequences in the Pandit database

way too many!

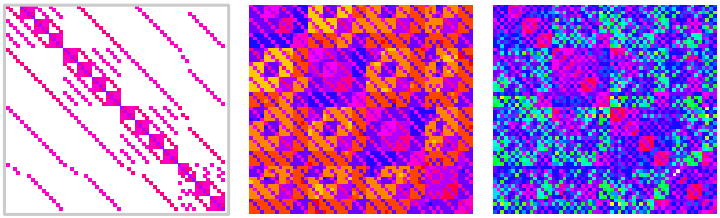
4. realism vs. complexity

codon rate matrix



4. realism vs. complexity

trying to find a balance



parameters :	2 (fitted)	23 (fitted)	1830 (not fitted)
lnL (Lysin) :	-4124.58	-3891.54	-4099.56

question: does this impact inference of positive selection?

preliminary answer: sometimes (but, oftentimes *no*)

4. realism vs. complexity

modeling amino acid exchangeabilities

multiple amino acid exhangeabilities:	model type
Yang and Goldman 1994; Yang et al. 1998;	PCP
Sanudiin et al. 2005; Wong et al. 2006	PCP
Conant and Stadler (2009)	PCP
Kosiol et al. 2007; De Maio et al. 2012	ECM ➔ MEP *
Doron-Faigenboin & Pupko 2007	MEP *
Miyazawa (2011)	MEP (LCEP) *
Zoller and Schneider (2012)	MEP (LCEP) *
Delport et al. 2010	GPP (GA)
Zaheri, Dib and Salamin (2014)	GPP *
Bielawski et al. (in prep.)	GPP *

PCP: physiochemcially constrained parametric model
MEP: mixed empirical and parametric model
ECM: empirical codon model
GPP: general purpose parametric model
* models permitting double and triple changes among codons

codon models

lots of options …

- methods have different pros and cons (that's OK)
- researchers have different preferences (that's OK)
- I use a wide variety of different methods; you might have different preferences (that's OK)
- most important: KNOW YOUR DATA and MAKE INFORMED DECISIONS.

no single method will be suitable for all situations