Comparison of synonymous and nonsynonymous rates to detect selection in protein-coding genes

Ziheng Yang
Department of Biology
University College London

Synonymous & nonsynonymous substitutions

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

Model of codon substitution

Factors to consider:

- Transition/transversion rate ratio: κ
- Biased codon usage: π_j for codon j
- Nonsynonymous/synonymous rate ratio ω

Matrix of relative rates: Q = $\{q_{ij}\}$

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

$$P(t) = e^{Qt}$$

(Goldman & Yang 1994 *Mol Biol Evol* **11**:725–736
Muse & Gaut 1994 *Mol Biol Evol* **11**:715–724)

Relative rates to CTG

Synonymous

CTC (Leu) → CTG (Leu) π_{CTG}
TTG (Leu) → CTG (Leu) $\kappa\pi_{CTG}$

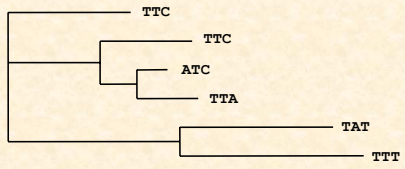
Nonsynonymous

GTG (Val) → CTG (Leu) $\omega\pi_{CTG}$
CCG (Pro) → CTG (Leu) $\kappa\omega\pi_{CTG}$

$\omega = d_N/d_S$ or K_A/K_S measures selection at the protein level

- $\omega = 1$: neutral evolution
- $\omega < 1$: negative (purifying) selection
- $\omega > 1$: positive (diversifying) selection

Likelihood calculation sums over all possible codons for each ancestral node

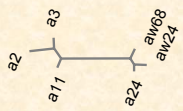


Extensions to basic model

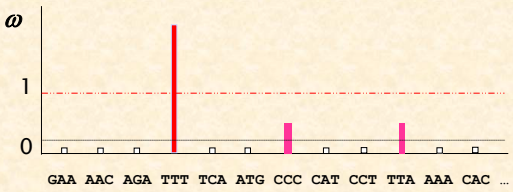
- Allow ω to differ among branches
- Allow ω to vary among sites
- Allow ω to vary both among branches and among sites

Data & information

a2	GGC	TCT	CAC	TCC	ATG	AGG	TAT	TTC	TTC	ACA	TCC
a24	...	..C	...	...	...	..T	...	...	.A.	..C	...
a11	...	..C	..A	...	...	...	...	...	.A.	..C	...
aw24	...	..C	...	...	...	...	...	...	CA.	..C	...
aw68	...	..C	...	...	...	..A	...	...	.A.	..C	...
a3	...	..T	..T	...	...	...	...	...	C..	..T	...



Variable selective pressures among sites



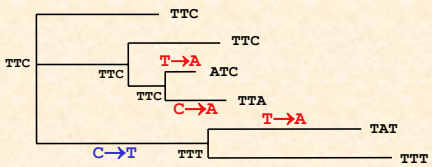
Two questions

- Are there any sites under positive selection with $\omega > 1$?
- Where are those sites?

Possible approaches

- Test each site for positive selection (Suzuki & Gojobori 1999 *Mol. Biol. Evol.* 16: 1315–1328)
- Decide on which sites might be under selection and focus on them (Hughes & Nei 1988 *Nature* 335:167–170) (fixed-sites model)
- Use a statistical distribution to model the ω variation (random-sites model, fishing expedition)

A simple approach



3 nonsynonymous changes
1 synonymous change

Drawbacks

- Different rates of change between nucleotides (transition/transversion rate bias and codon usage bias)
- Uncertainties in ancestral reconstruction
- Branch lengths and multiple hits
- Doing statistics with one data point

LRT for sites under positive selection

H_0: there are no sites at which $\omega > 1$
H_1: there are some such sites

Compare $2\Delta\ell = 2(\ell_1 - \ell_0)$ with a χ^2 distribution

Nielsen & Yang 1998 Genetics **148**:929–936
Yang, et al. 2000. Genetics **155**:431–449

Table 2

Model code	p	Parameters	Notes
M0 (one-ratio)	1	ω	One ω ratio for all sites
M1 (neutral)	1	p_0	$p_1 = 1 - p_0$, $\omega_0 = 0$, $\omega_1 = 1$
M2 (selection)	3	p_0, p_1, ω_2	$p_2 = 1 - p_0 - p_1$, $\omega_0 = 0$, $\omega_1 = 1$
M3 (discrete)	$2K - 1$ ($K = 3$)	$p_0, p_1, \ldots, p_{K-2}$ $\omega_0, \omega_1, \ldots, \omega_{K-1}$	$p_{K-1} = 1 - p_0 - p_1 - \cdots - p_{K-2}$
M4 (freqs)	$K - 1$ ($K = 5$)	$p_0, p_1, \ldots, p_{K-2}$	The ω_k are fixed at 0, $\frac{1}{3}$, $\frac{2}{3}$, 1, and 3
M5 (gamma)	2	α, β	from $\Gamma(\alpha, \beta)$
M6 (2gamma)	4	$p_0, \alpha_0, \beta_0, \alpha_1$	p_0 from $\Gamma(\alpha_0, \beta_0)$ and $p_1 = 1 - p_0$ from $\Gamma(\alpha_1, \alpha_1)$
M7 (beta)	2	p, q	from $B(p, q)$
M8 (beta&ω)	4	p_0, p, q, ω	p_0 from $B(p, q)$ and $1 - p_0$ with ω
M9 (beta&gamma)	5	p_0, p, q, α, β	p_0 from $B(p, q)$ and $1 - p_0$ from $\Gamma(\alpha, \beta)$
M10 (beta&gamma+1)	5	p_0, p, q, α, β	p_0 from $B(p, q)$ and $1 - p_0$ from $1 + \Gamma(\alpha, \beta)$
M11 (beta&normal>1)	5	p_0, p, q, μ, σ	p_0 from $B(p, q)$ and $1 - p_0$ from $N(\mu, \sigma^2)$, truncated to $\omega > 1$
M12 (0&2normal>1)	5	$p_0, p_1, \mu_2, \sigma_1, \sigma_2$	p_0 with $\omega_0 = 0$, and $1 - p_0$ from the mixture: p_1 from $N(1, \sigma_1^2)$, and $1 - p_1$ from $N(\mu_2, \sigma_2^2)$, both Normals truncated to $\omega > 1$
M13 (3normal>0)	6	$p_0, p_1, \mu_2, \sigma_0, \sigma_1, \sigma_2$	p_0 from $N(0, \sigma_0^2)$, p_1 from $N(1, \sigma_1^2)$, and $p_2 = 1 - p_0 - p_1$ from $N(\mu_2, \sigma_2^2)$, all Normals truncated to $\omega > 1$

Two pairs of useful models

M1a (Nearly Neutral)

Site class k:	0	1
p_k:	p_0	p_1
ω_k:	$\omega_0 < 1$	$\omega_1 = 1$

M2a (Positive Selection)

Site class k:	0	1	2
p_k:	p_0	p_1	p_2
ω_k:	$\omega_0 < 1$	$\omega_1 = 1$	$\omega_1 > 1$

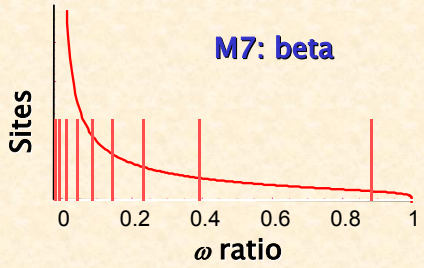
M7 (beta, using 10 site classes)

$\omega \sim$ beta(p, q)

M8 (beta&ω)

p_0 of sites from beta(p, q)
$p_1 = 1 - p_0$ of sites with $\omega_s > 1$

Discretisation of a continuous distribution



Likelihood for estimating parameters η in the ω distribution

$$f(x_j;\eta) = \int f(x_j \mid \omega,\eta) f(\omega \mid \eta) d\omega$$
$$= \sum_k p_k f(x_j \mid \omega_k,\eta)$$

Empirical Bayes for estimating ω for site

$$f(\omega_k \mid x_j,\eta) = p_k f(x_j \mid \omega_k,\eta) / f(x_j,\eta)$$

Human MHC Class I data, including loci A, B, C

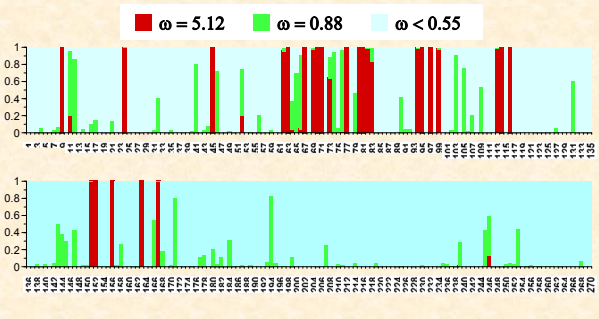
192 alleles, 270 codons

Likelihood values and parameter estimates for Class I MHC alleles

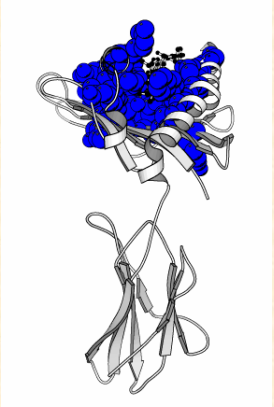
Model	ℓ	Parameter estimates
M7 (beta)	–7,498.97	beta(0.10, 0.35)
M8 (beta&ω)	–7,232.68	p_0 = 0.90, beta(0.17, 0.71)
		(p_1 = 0.10), ω = 5.12

Likelihood ratio test of positive selection:
$2\Delta\ell = 2 \times 266.29 = 532.58$, $P < 0.000$, d.f. = 2

Posterior probabilities for MHC



25 sites identified by M8 (beta&ω)



Fixed-sites models for partitioned data

57 ARS sites:

5M, 7Y, 9F, 22F, 24A, 26G, 57P, 58E, 59Y, 61D, 62G, 63E, 64T, 65R, 66K, 67V, 68K, 69A, 70H, 71S, 72Q, 73T, 74H, 75R, 76V, 77D, 80T, 81L, 82R, 84Y, 95V, 97R, 99Y, 114H, 116Y, 143T, 145H, 146K, 147W, 149A, 150A, 151H, 152E, 154E, 155Q, 156Q, 157R, 158A, 159Y, 161E, 162G, 163T, 165V, 166E, 167W, 169R, and 171Y.

213 non-ARS sites.

Bjorkman, et al. 1987. Nature 329:512-518
Hughes & Nei. 1988. Nature 335:167-170

Fixed-sites model for partitioned data for Class I MHC

Model	ℓ	Parameter estimates
Two partitions	-7,671.92	$\omega_{non-ARS} = 0.23$, $\omega_{ARS} = 1.86$
Two partitions	-7,681.25	$\omega_{non-ARS} = 0.23$, $\omega_{ARS} = 1$ fixed

$2\Delta\ell = 18.66^{**}$, d.f. = 1

Comparison between fixed-and random-sites models

Model	ℓ	Parameter estimates
Two partitions	-7,671.92	$\omega_{non-ARS} = 0.23$, $\omega_{ARS} = 1.86$
M8 (beta&ω)	-7,232.68	$p_0 = 0.90$, beta(0.17, 0.71) ($p_1 = 0.10$), $\omega = 5.12$

Comparison between fixed-and random-sites models

- 22 of the 25 sites identified by M8 are in the ARS list. 3 sites (45M, 94T, and 113Y) are in the ARS domain.
- Fixed-sites models fit the data more poorly because the 57 sites at ARS included highly conserved sites.

Advantages of ML

- Accounts for the genetic code
- Accounts for ts/tv rate bias and codon usage bias
- Avoids bias in ancestral sequence reconstruction
- Uses probability theory to correct for multiple hits

Rate matrix Q = $\{q_{ij}\}$

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

Limitations

- Same selective pressure for all lineages
- No recombination within the sequence
- No variation in synonymous rate among sites
- Same rate for all amino acid changes
- One ω for positive selection sites
- No sequencing or alignment errors

Limitations

- Posterior probability calculation (naïve empirical Bayes) does not account for sampling errors in parameter estimates.
- The level of sequence divergence and the number of sequences are two major factors affecting accuracy and power. Data of only a few closely related sequences do not contain much information.

Acknowledgments

Nick Goldman
Rasmus Nielsen
Anne-Mette Pedersen
Willie Swanson
Maria Anisimova

BBSRC

http://abacus.gene.ucl.ac.uk/