Codon models and positive selection in protein evolution

Ziheng Yang

Department of Biology
University College London

Plan

- Positive selection & its importance
- Methods for detecting positive selection
- Detecting amino acid sites under positive selection
- Genes detected to be under positive selection

There are two main explanations for genetic variation observed within a population or between species:

Natural selection (survival of the fittest)
Mutation and drift (survival of the luckiest)

Gillespie, J.H. 1998. *Population genetics: a concise guide.* John Hopkins University Press, Baltimore.

Hartl, D.L., and A.G. Clark. 1997. *Principles of population genetics.* Sinauer Associates, Sunderland, Massachusetts.

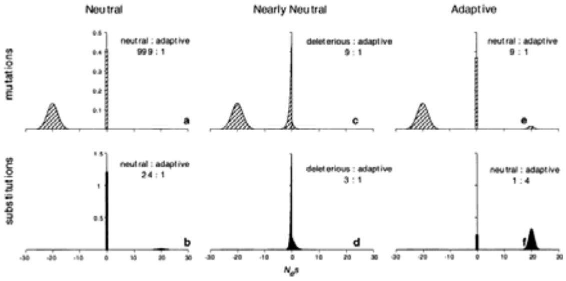
Positive & negative selection

Genotype	AA	Aa	aa
Frequency	p^2	$2p(1-p)$	$(1-p)^2$
Fitness	1	$1+s$	$1+2s$

(A: “wild-type allele”; a: new mutant)

s is selection coefficient:
$s \approx 0$: neutral evolution
$s < 0$: negative (purifying) selection
$s > 0$: positive selection (adaptive evolution)

Theories of molecular evolution



Akashi, H. (1999) *Gene* 238: 39–51

Detecting selection is useful

- **for testing evolutionary theory**
- **for identifying functional elements in genomes.**

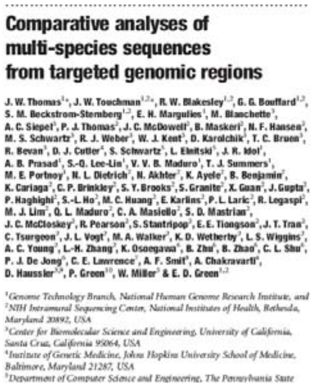
Evolutionary conservation means function

Genes or genome regions conserved across diverse species most likely have some functional significance.

Conservation → function

About 12Mb of the cystic fibrosis region were sequenced in 12 vertebrate and fish species, and used to identify a number of conserved non-coding segments previously unknown. Closely related mammalian species are effective in identifying regulatory elements while distantly related species are effective in identifying coding regions.

(Thomas, et al. 2003. *Nature* 424:788-793)



Comparative analyses of multi-species sequences from targeted genomic regions

J. W. Thomas[1*], J. W. Touchman[1,2*], R. W. Blakesley[1,2], G. G. Bouffard[1,2], S. M. Beckstrom-Sternberg[1,2], E. H. Margulies[1], M. Blanchette[3], A. C. Siepel[3], P. J. Thomas[2], J. C. McDowell[2], B. Maskeri[2], N. F. Hansen[2], M. S. Schwartz[3], R. J. Weber[3], W. J. Kent[3], D. Karolchik[3], T. C. Bruen[3], R. Bevan[3], D. J. Cutler[4], S. Schwartz[5], L. Elnitski[5], J. R. Idol[1], A. B. Prasad[1], S.-Q. Lee-Lin[1], V. V. B. Maduro[1], T. J. Summers[1], M. E. Portnoy[1], N. L. Dietrich[2], N. Akhter[2], K. Ayele[2], B. Benjamin[2], K. Cariaga[2], C. P. Brinkley[2], S. Y. Brooks[2], S. Granite[2], X. Guan[2], J. Gupta[2], P. Haghighi[2], S.-L. Ho[2], M. C. Huang[2], E. Karlins[2], P. L. Laric[2], R. Legaspi[2], M. J. Lim[2], Q. L. Maduro[2], C. A. Masiello[2], S. D. Mastrian[2], J. C. McCloskey[2], R. Pearson[2], S. Stantripop[2], E. E. Tiongson[2], J. T. Tran[2], C. Tsurgeon[2], J. L. Vogt[2], M. A. Walker[2], K. D. Wetherby[2], L. S. Wiggins[2], A. C. Young[2], L.-H. Zhang[2], K. Osoegawa[6], B. Zhu[6], B. Zhao[6], C. L. Shu[6], P. J. De Jong[6], C. E. Lawrence[7], A. F. Smit[8], A. Chakravarti[4], D. Haussler[3,9], P. Green[10], W. Miller[5] & E. D. Green[1,2]

[1] *Genome Technology Branch, National Human Genome Research Institute, and* [2] *NIH Intramural Sequencing Center, National Institutes of Health, Bethesda, Maryland 20892, USA*
[3] *Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA*
[4] *Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21287, USA*
[5] *Department of Computer Science and Engineering, The Pennsylvania State*

High variability may also mean functional significance, if the variability is driven by selection.

Evolutionary biologists are more interested in positive selection because fixations of advantageous mutations in the genes or genomes are responsible for evolutionary innovations and species divergences.

Positive selection can be detected using population genetics tests of neutrality

- McDonald & Kreitman test (1991)
- Hudson, Kreitman and Aquade (HKA) test (1987)
- Fu & Li test (1993)
- Fay, Wyckoff & Wu (2002)

Fay JC, Wu CI. 2003. *Annu. Rev. Genomics. Hum. Genet.* 4:213-235.
Kreitman, M. 2000. *Annu. Rev. Genomics Hum. Genet.* 1:539-559.
Nielsen R. 2005. *Annu. Rev. Genet* 39:197-218.

Positive selection can also be detected through phylogenetic comparison of synonymous and nonsynonymous substitution rates

- **$\omega = 1$: neutral evolution ($s = 0$)**
- **$\omega < 1$: negative (purifying) selection ($s < 0$)**
- **$\omega > 1$: positive (diversifying) selection ($s > 0$)**

(Miyata and Yasunaga 1980; Gojobori 1983;
Li *et al.* 1985; Nei & Gojobori 1986)

The nonsynonymous/synonymous rate ratio ω contrasts our expectations based on the genetic code and our observations after the filtering of selection on the protein.

If we expect N:S to be 74.5%:25.5% before selection on the protein, and observe 5:5 substitutions (differences), then

$$\omega = d_N/d_S = (5/5)/(74.5\%/25.5\%) = 0.34$$

Definitions

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

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

$\omega = d_N / d_S$: nonsynonymous/synonymous rate ratio

Codon-substitution model: 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) $\kappa\omega\pi_{CTG}$

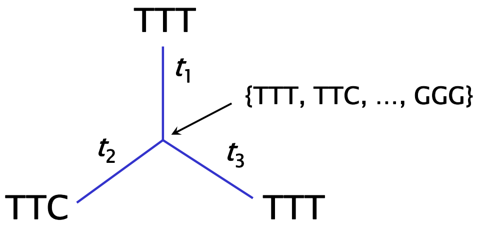
Rate matrix $Q = \{q_{ij}\}$

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

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

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

Likelihood calculation on a tree sums over all possible codons for each ancestral node



Codon substitution models

- ***Branch models*** **to test positive selection on lineages on the tree**
  (Yang 1998. *Mol. Biol. Evol.* 15:568-573)
- ***Site models*** **to test positive selection affecting individual sites**
  (Nielsen & Yang. 1998. *Genetics* 148:929-936; Yang, *et al.* 2000. *Genetics* 155:431-449)
- ***Branch-site models*** **to detect positive selection at a few sites on a particular lineage**
  (Yang & Nielsen. 2002. *Mol. Biol. Evol.* 19:908-917; Yang, *et al.* 2005. *Mol. Biol. Evol.* 22:1107-1118)

Branch models

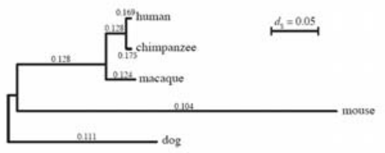
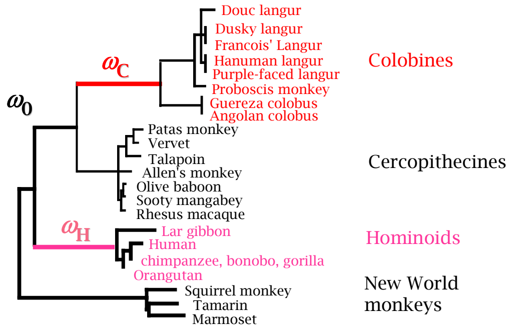


Figure S6.2: An estimate of ω for each branch of a five-species phylogeny. Show is the maximum-likelihood phylogeny for 5286 orthologous quintents, with branch lengths drawn in proportion to the estimated number of synonymous substitutions per synonymous site (d_S). Each branch is labeled with the corresponding estimate of ω.

Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the Rhesus macaque genome. *Science* 316:222-234.

Adaptive evolution in primate lysozyme



Log-likelihood values and parameter estimates

Model	p	ℓ	ω_0	ω_C
A. 1-ratio: ω_0=ω_C	35	−1043.84	0.574	= ω_0
B. 2-ratios: ω_0, ω_C	36	−1041.70	0.489	3.383
C. 2-ratios: ω_0, ω_C=1	35	−1042.50	0.488	1

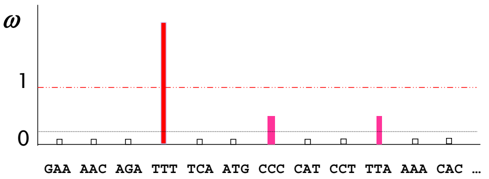
(Yang 1998 *Mol. Biol. Evol.* **15**: 568-573
Data from Messier & Stewart 1997 *Nature* **385**: 151-154)

Likelihood ratio test statistics

Null hypothesis	$2\Delta\ell$
ω_C = ω_0	4.24*
ω_C = 1	1.60

Site models

Early studies average synonymous and nonsynonymous rates over sites and have little power in detecting adaptive evolution.



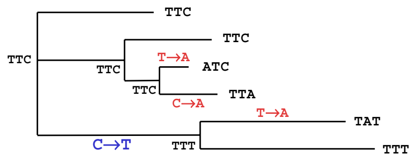
Possible approaches

- **Estimate and test one ω for every site**
  (Fitch *et al.* 1997 PNAS 94:7712-7718; Suzuki & Gojobori 1999 *Mol. Biol. Evol.* 16: 1315-1328; Suzuki 2004 *J. Mol. Evol.* 59: 11-19; Massingham and Goldman 2005 *Genetics* 169: 1753-1762; Kosakovsky Pond and Frost 2005 *Mol. Biol. Evol.* 22: 1208-1222)

- **Focus on sites potentially under selection based on structure**
  (Hughes & Nei 1988 *Nature* 335:167-170; Yang & Swanson 2002 Mol. Biol. Evol. 19: 49-57) (fixed-sites model)

- **Use a statistical distribution to model the ω variation**
  (Nielsen & Yang 1998 *Genetics* 148: 929-936; Yang *et al.* 2000 *Genetics* 155: 431-449) (random-sites model, fishing expedition)

one ω for every site



3 nonsynonymous changes
1 synonymous change

The approach of one ω for a site uses too many parameters.

The standard approach to dealing with the problem is to assign a prior on ω and use a nonparametric or parametric empirical Bayes approach.

Use of codon models to detect amino acid sites under diversifying selection

- **Likelihood ratio test (LRT) for sites under positive selection**
- **Empirical Bayesian calculation of posterior probabilities of sites under positive selection**

LRT of sites under positive selection

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

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

(Nielsen & Yang 1998 Genetics **148**:929–936;
Yang, Nielsen, Goldman & Pedersen 2000. Genetics **155**:431–449)

Two pairs of useful models

M1a (neutral)

Site class:	**0**	**1**
Proportion:	p_0	p_1
ω **ratio:**	$\omega_0<1$	$\omega_1=1$

M2a (selection)

Site class:	**0**	**1**	**2**
Proportion:	p_0	p_1	p_2
ω **ratio:**	$\omega_0<1$	$\omega_1=1$	$\omega_2>1$

Modified from Nielsen & Yang (1998), where ω_0=0 is fixed

M7 (beta)

ω ~ **beta(p, q)**

M8 (beta&ω)

p_0 **of sites from beta(p, q)**

$p_1 = 1 - p_0$ **of sites with $\omega_s > 1$**

Yang, Nielsen, Goldman, Pedersen (2000 Genetics **155**:431-449)

Human MHC Class I data: 192 alleles, 270 codons

Model	ℓ	Parameter estimates
M1a (neutral)	−7,490.99	$p_0 = 0.830$, $\omega_0 = 0.041$ $p_1 = 0.170$, $\omega_1 = 1$
M2a (selection)	−7,231.15	$p_0 = 0.776$, $\omega_0 = 0.058$ $p_1 = 0.140$, $\omega_1 = 1$ $p_2 = 0.084$, $\omega_2 = 5.389$

Likelihood ratio test of positive selection:
$2\Delta\ell = 2 \times 259.84 = 519.68$, $P < 0.000$, d.f. = 2

Empirical Bayesian calculation of posterior probabilities that a site is under positive selection with $\omega > 1$.

- Naïve Empirical Bayes (NEB) ignores sampling errors in parameter estimates.
- Bayes Empirical Bayes (BEB) accounts for sampling errors by integrating over a prior.

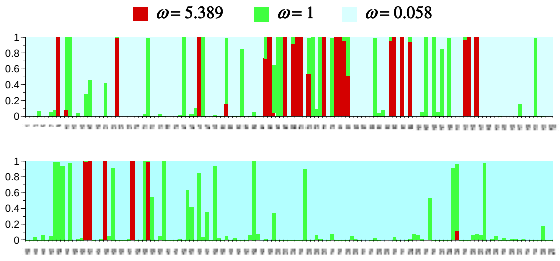
Nielsen & Yang. 1998. *Genetics* **148**:929-936.
Yang, Wong & Nielsen. 2005. *Mol. Biol. Evol.* **22**:1107-1118.

Naïve Empirical Bayes (NEB)

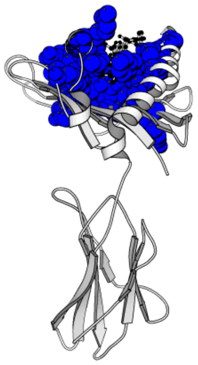
Under M2a, there are
Three site classes: $\omega_0 = 0.058$, $\omega_1 = 1$, $\omega_2 = 5.389$
Prior proportions: $p_0 = 0.776$, $p_1 = 0.140$, $p_2 = 0.084$

Bayes's theorem is used to calculate the posterior probabilities for the three site classes for each site, given the data.

Posterior probabilities for MHC (M2a)



25 sites identified under M2a



With more genomes sequenced, the approach of evolutionary comparison will become more powerful. It provides a way of generating interesting biological hypotheses, which can be validated by experimentation.

Ivarsson, Y., A. J. Mackey, M. Edalat, W. R. Pearson, and B. Mannervik. 2002. Identification of residues in glutathione transferase capable of driving functional diversifcation in evolution: a novel approach to protein design. *J. Biol. Chem.* 278:8733-8738.

Bielawski, J. P., K. A. Dunn, G. Sabehi, and O. Beja. 2004. Darwinian adaptation of proteorhodopsin to different light intensities in the marine environment. *Proc. Natl. Acad. Sci. U.S.A.* 101:14824-14829.

Sawyer, S. L., L. I. Wu, M. Emerman, and H. S. Malik. 2005. Positive selection of primate TRIM5á identifies a critical species-specific retroviral restriction domain. Proc. Natl. Acad. Sci. U.S.A. 102:2832-2837.

Advantages of ML

- Accounts for the genetic code
- Accounts for transition-transversion rate differences and codon usage
- Avoids bias in ancestral reconstruction
- Uses probability theory to correct for multiple hits

Disadvantages of ML

- Model assumptions may be unrealistic.
- The method detects positive selection only if it generates excessive nonsynonymous substitutions. It may lack power in detecting one-off directional selection or when the sequences are highly similar or highly divergent. It is typically useless for population data.

Which proteins are under positive selection?

- Host proteins involved in defence or immunity against viral, bacterial, fungal or parasite attacks (MHC, immunoglobulin VH, class 1 chitinas).
- Viral or pathogen proteins involved in evading host defence (HIV env, nef, gap, pol, etc., capsid in FMD virus, flu virus hemagglutinin gene).
- Proteins or pheromones involved in reproduction (abalone sperm lysin, sea urchin bindin, proteins in mammals).
- Proteins that acquired new functions after gene duplication.
- Miscellaneous (diet, globins,).

Further reading

Fay JC, Wu CI. 2003. Sequence divergence, functional constraint, and selection in protein evolution. *Annu. Rev. Genomics. Hum. Genet.* 4:213-235.

Nielsen R. 2005. Molecular signatures of natural selection. *Annu. Rev. Genet* 39:197-218.

Yang Z. 2002. Inference of selection from multiple species alignments. *Curr. Opinion Genet. Devel.* **12**:688-694.

Yang Z. 2006. *Computational Molecular Evolution.* OUP, Chapter 8