

# Comparison between two protein-coding DNA sequences

**Ziheng Yang**

Department of Biology  
University College London

## Outline

- Estimation of  $d_N$  and  $d_S$  between two sequences
  - Counting methods
  - Codon substitution model
  - ML method
- Analysis of multiple sequences on a tree

## Synonymous (silent) and nonsynonymous (replacement) substitutions

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

## 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

## Counting method

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

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

(2) The gene is 3×300 nucleotides long, so

$$S = 900 \times 25.5\% = 229.5$$

$$N = 900 \times 74.5\% = 670.5$$

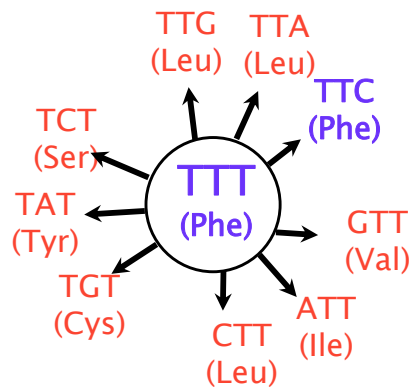
$$d_S = 5/229.5 = 0.0218$$

$$d_N = 5/670.5 = 0.0075$$

## Counting method

- Count the numbers of synonymous and nonsynonymous sites ( $S$  and  $N$ )
- Count the numbers of synonymous and nonsynonymous differences ( $S_d$  and  $N_d$ )
- Calculate the proportions of different sites and correct for multiple hits

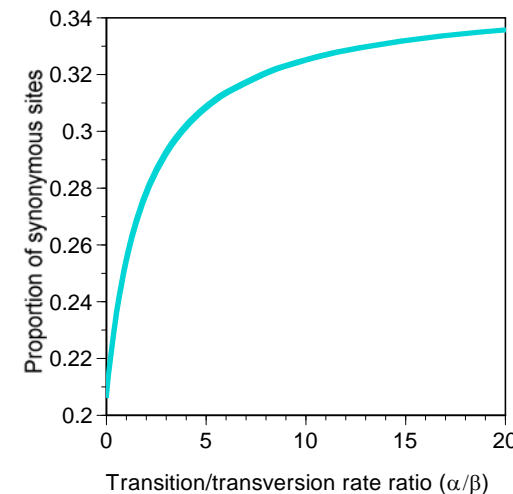
## Counting sites ( $S$ and $N$ )



Codon TTT has  
1/3 synonymous sites  
8/3 nonsynonymous sites

Sites are defined as mutational opportunities

## The impact of transition-transversion rate difference



At the third position, transitions are more likely to be synonymous than transversions.

## Codon usage bias

Analysis of real genes suggests that codon usage bias leads to reduced number of synonymous sites (has the opposite effect to the ts/tv bias).

## Counting differences

Two pathways between CCT and CAG:

Pathway	Syn	Nonsyn
CCT (Pro) ↔ CAT (His) ↔ CAG (Gln)	0	2
CCT (Pro) ↔ CCG (Pro) ↔ CAG (Gln)	1	1
Average	0.5	1.5

## Correcting for multiple hits

*Ad hoc* correction using nucleotide-based models, which assume that a nonsynonymous site has equal rate of changing into 3 other nonsynonymous nucleotides (Lewontin 1989).

## Interpretation

$d_S$  is the expected number of nucleotide substitutions per nucleotide site, averaged over 3 codon positions, before selection on the protein (i.e., had there be no selection on the protein)

$$d_N = d_S \times \omega$$

## Counting method

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.

## Human and orangutan $\alpha 2$ -globin genes (142 codons)

Method/Model	$\kappa$	$S$	$N$	$d_N$	$d_S$	$d_N/d_S$
NG86	1	109.4	316.6	0.0095	0.0569	0.168
Ina95	2.1	119.3	299.9	0.0101	0.0523	0.193
YN00	6.1	61.7	367.3	0.0083	0.1065	0.078

Base frequencies at 3rd position:  
 T = 9%, C = 52%, A = 1%, G = 37%  
 (Yang & Bielawski 2000. *TREE* 15:496–503)

## Markov chain model of codon substitution

### Factors to consider:

- Transition/transversion rate ratio:  $\kappa$
- Biased codon usage:  $\pi_j$  for codon  $j$
- Nonsynonymous/synonymous rate ratio:  $\omega = d_N/d_S$

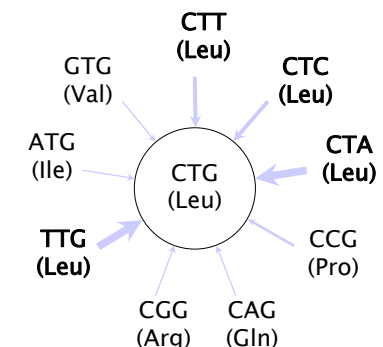
## Codon-substitution model: Rates to CTG

### Synonymous

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

### Nonsynonymous

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



## 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)

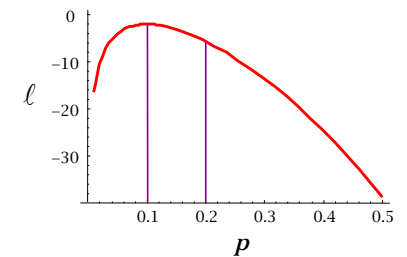
## Likelihood is the probability of the data, viewed as a function of the unknown parameters

**Example.** There are many red and blue fish in a pond. We want to estimate the proportion of red fish in the pond ( $p$ ). We take a sample of  $n = 100$  fish and found  $x = 10$  red and  $n - x = 90$  blue.

$$L(p; x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{100}{10} p^{10} (1-p)^{90}$$

$$\ell(p; x) = \log \binom{100}{10} + 10 \log(p) + 90 \log(1-p)$$

$$\hat{p} = 10/100 = 0.1$$



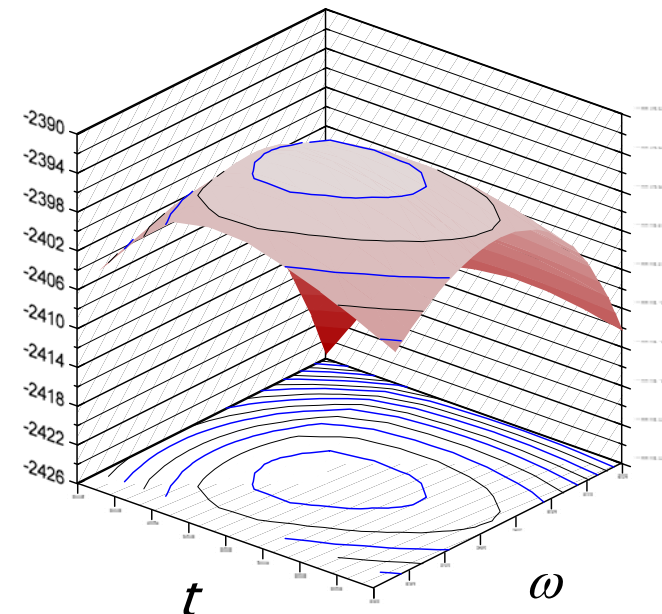
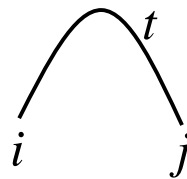
## ML estimation of $d_S$ and $d_N$

The probability of observing a site with codons  $i$  and  $j$  in the two sequences is

$$\pi_i p_{ij}(t)$$

The log likelihood is

$$\ell(t, \kappa, \omega) = \sum_{h=1}^n \log \{ \pi_i p_{ij}(t) \}$$



MLEs of  $t$   
 and  $\omega \rightarrow$   
 MLEs of  $d_S$   
 and  $d_N$   
 ( $\kappa=1$  fixed)

## ML estimation of $d_S$ and $d_N$

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

## Chapman–Kolmogorov theorem

$$p_{ij}(s+t) = \sum_k p_{ik}(s) p_{kj}(t)$$

$i = \text{TTT}$   
 $\downarrow s$   
 $k = \{\text{TTT}, \text{TTC}, \dots, \text{GGG}\}$   
 $\downarrow t$   
 $j = \text{TTT}$

Given codon  $i$  now, the probability that it will be  $j$  time ( $s+t$ ) later is a sum over all possible states at any intermediate time  $s$ .

## Time reversibility

Almost all models used in molecular phylogenetics are time reversible. The Markov chain is said to be *time reversible* if and only if

$$\pi_i q_{ij} = \pi_j q_{ji}, \text{ for all } i \neq j.$$

which is the same requirement as

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t), \text{ for all } i \neq j.$$

## Reversibility means no root

$$\Pr(ij | t_1, t_2)$$

$$= \sum_k \pi_k p_{ki}(t_1) p_{kj}(t_2)$$

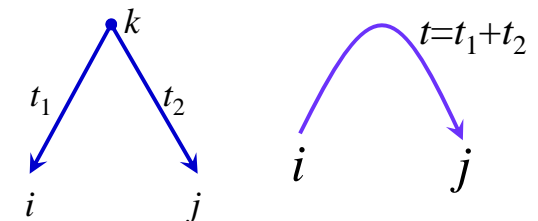
← reversibility

$$= \sum_k \pi_i p_{ik}(t_1) p_{kj}(t_2)$$

$$= \pi_i \sum_k p_{ik}(t_1) p_{kj}(t_2)$$

← Chapman-Kolmogorov theorem

$$= \pi_i p_{ij}(t_1 + t_2)$$



## Human and orangutan $\alpha_2$ -globin genes (142 codons)

Method/Model	$\kappa$	$S$	$N$	$d_N$	$d_S$	$d_N/d_S$
NG86	1	109.4	316.6	0.0095	0.0569	0.168
Ina95	2.1	119.3	299.9	0.0101	0.0523	0.193
YN00	6.1	61.7	367.3	0.0083	0.1065	0.078
<b>ML (GY94)</b>						
(1) ML Fequal, $\kappa = 1$	1	108.5	317.5	0.0093	0.0557	0.167
(2) ML Fequal, $\kappa$ estimated	3.0	124.6	301.4	0.0099	0.0480	0.206
(7) ML F61, $\kappa = 1$ fixed	1	58.3	367.7	0.0082	0.1145	0.072
(8) ML F61, $\kappa$ estimated	5.3	55.3	370.7	0.0082	0.1237	0.066

Base frequencies at 3rd position:  
 T = 9%, C = 52%, A = 1%, G = 37%  
 (Yang & Bielawski 2000. *TREE* 15:496-503)

## Comments on methods

- Assumptions matter more than methods.
- Ignoring the transition/transversion rate bias leads to underestimation of  $S$ , overestimation of  $d_S$  and underestimation of the  $\omega$  ratio.
- Codon-usage bias often has the opposite effect to the transition/transversion bias and can be more important.
- Different methods can produce different estimates even when the sequences are highly similar.
- The counting methods are safe to use if codon usage (especially at the 3rd position) is uniform, the sequences are not very divergent, and transition and transversion rates are similar (NG86).

## Common problems in estimating $d_S$ and $d_N$ in comparative genomics

1. The most common problem is wrong sequence divergence, often too high, but sometimes too low.
  - $d_S > 1$ : large sampling errors.
  - $d_S > 5$ : unreliable.
2. Data quality control, alignment.

## Software

Methods	Software
<b>Counting methods</b>	<b>MEGA; codeml &amp; yn00 in PAML</b>
<b>NG86</b>	<b>PAML</b>
<b>Li93</b>	<b>MEGA, DAMBE, codeml</b>
<b>Comeron 95</b>	<b>DIVERGE by Comeron</b>
<b>YN00</b>	<b>yn00 in PAML</b>
<b>ML methods</b>	
<b>GY94</b>	<b>codeml</b>

# Codon models and positive selection in protein evolution

**Ziheng Yang**

Department of Biology  
University College London

## Outline

- 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)

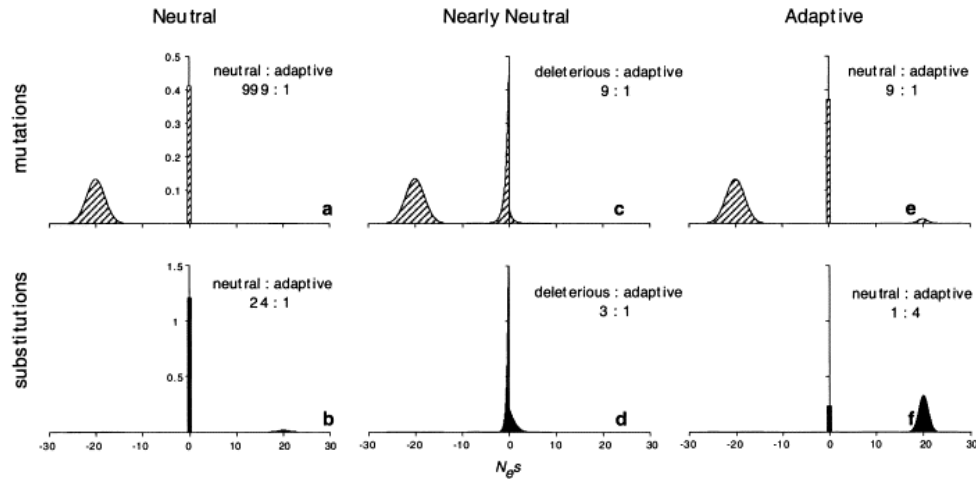
$s$  is selection coefficient:

$s \approx 0$ : neutral evolution

$s < 0$ : negative (purifying) selection

$s > 0$ : positive selection (adaptive evolution)

# Theories of molecular evolution



Akashi, H. (1999) *Gene* 238: 39-51

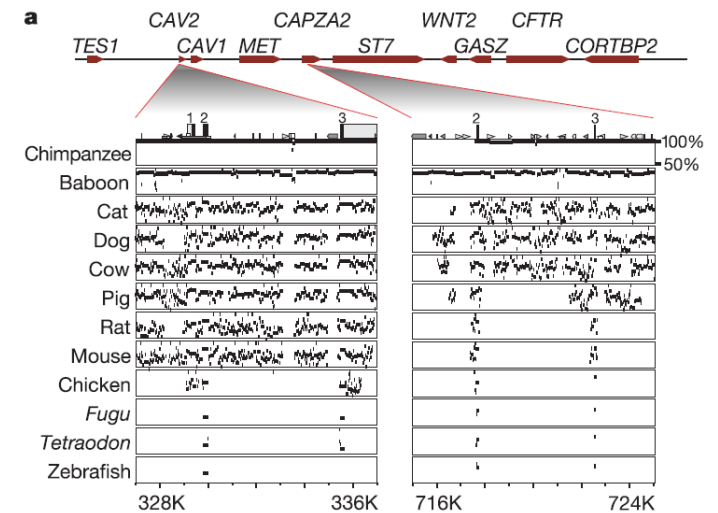
# Detecting selection is useful

- 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



Percentage identity when human is aligned with another species. Close species are effective in identifying regulatory elements while distant species are effective in identifying coding regions.

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

## 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 $\omega$ contrasts our expectations based on the genetic code and our observations after the filtering of selection on the protein.

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

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

## Codon-substitution model ( $Q_{61 \times 61}$ ): Rates to CTG

### Synonymous

CTC (Leu) → CTG (Leu)  $\pi_{\text{CTG}}$

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

### Nonsynonymous

GTC (Val) → CTG (Leu)  $\omega\pi_{\text{CTG}}$

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

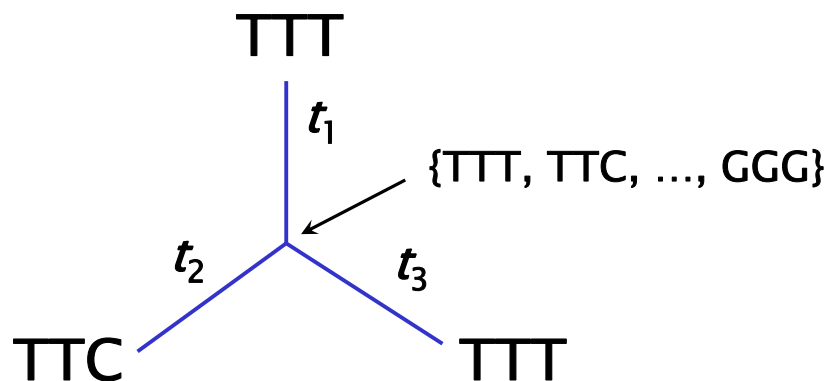
## Rate matrix $Q = \{q_{ij}\}$

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

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

(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



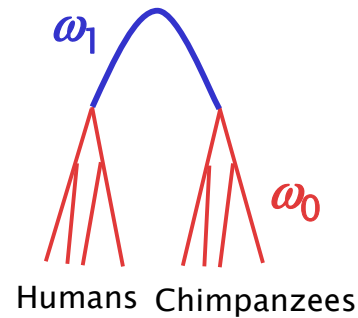
## Codon substitution models

- **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)

## McDonald–Kreitman test under codon models

### The LRT

- corrects for multiple hits
- is usable for multiple species



(Hasegawa, Cao & Yang 1998 MBE 15:1499–1505)

## Branch models

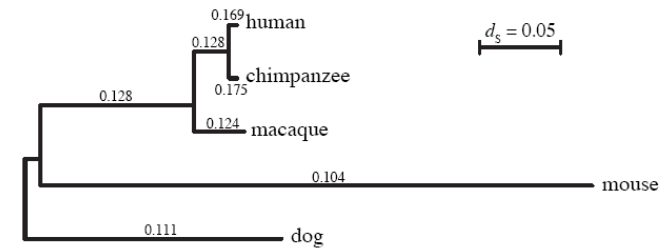


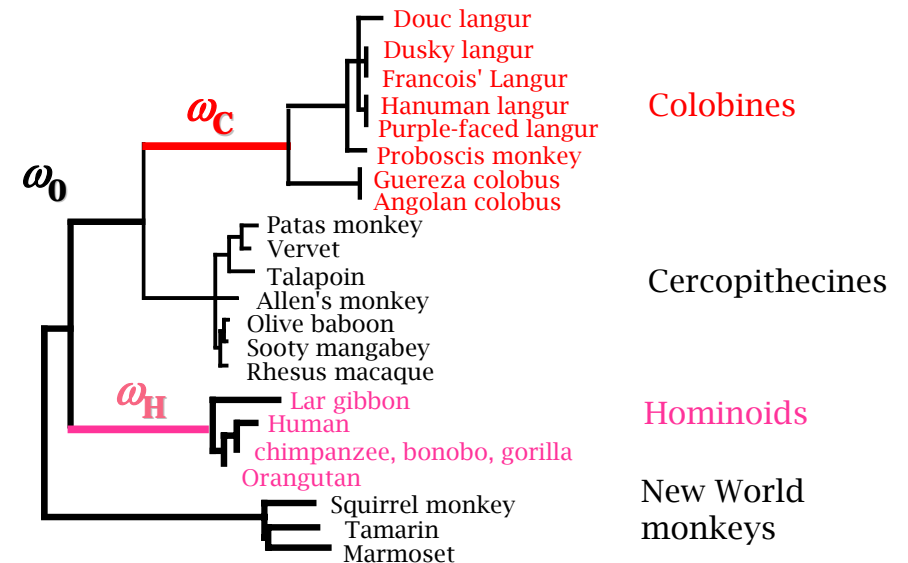
Figure S6.2: An estimate of  $\omega$  for each branch of a five-species phylogeny. Show is the maximum-likelihood phylogeny for 5286 orthologous quintets, with branch lengths drawn in proportion to the estimated number of synonymous substitutions per synonymous site ( $d_s$ ). Each branch is labeled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

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

## Likelihood ratio test to compare two nested models

If the more general (alternative) model  $H_1$  has  $p$  parameters with log likelihood  $\ell_1$ , and the simpler (null) model  $H_0$  has  $q$  parameters with log likelihood  $\ell_0$ . Then twice the log likelihood difference,  $2\Delta\ell = 2(\ell_1 - \ell_0)$ , can be compared with the  $\chi^2$  distribution with d.f. =  $p - q$  to test whether the simpler model is rejected.

## Adaptive evolution in primate lysozyme



## Log-likelihood values and parameter estimates

Model	$p$	$\ell$	$\omega_0$	$\omega_C$
A. 1- ratio: $\omega_0 = \omega_C$	35	-1043.84	0.574	$= \omega_0$
B. 2- ratios: $\omega_0, \omega_C$	36	-1041.70	0.489	3.383
C. 2- ratios: $\omega_0, \omega_C = 1$	35	-1042.50	0.488	1

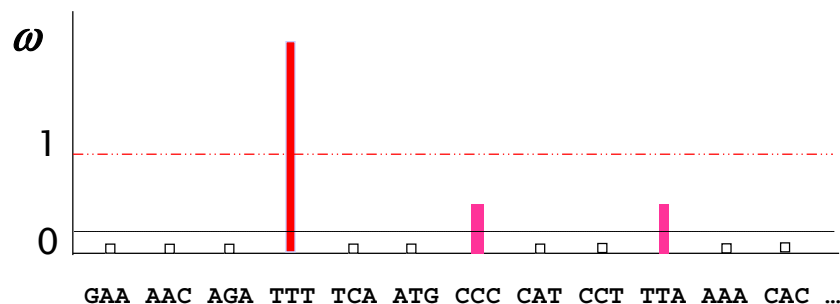
(Yang 1998 *Mol. Biol. Evol.* **15**: 568-573  
Data from Messier & Stewart 1997 *Nature* **385**: 151-154)

## Likelihood ratio test statistics

Null hypothesis	$2\Delta\ell$
$\omega_C = \omega_0$	4.24*
$\omega_C = 1$	1.60

## Site models

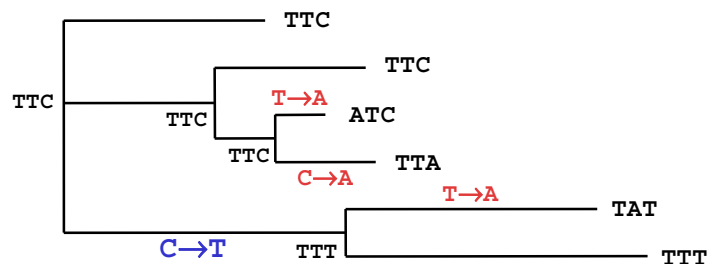
Early studies average synonymous and nonsynonymous rates over sites and have little power in detecting adaptive evolution.



## Possible approaches

- Estimate and test one  $\omega$  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  $\omega$  variation  
(Nielsen & Yang 1998 *Genetics* 148: 929–936; Yang *et al.* 2000 *Genetics* 155: 431–449) (**random-sites model, fishing expedition**)

## one $\omega$ for a site



**3 nonsynonymous changes**  
**1 synonymous change**

### ML programs

HyPhy (Kosakovsky Pond and Frost 2005 *Mol. Biol. Evol.* 22: 1208–1222)

SLR (Massingham and Goldman 2005 *Genetics* 169: 1753–1762)

The approach of one  $\omega$  for a site uses too many parameters.

The standard approach to dealing with the problem of infinitely many parameters is to assign a prior on  $\omega$  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  $\chi^2$  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:    $p_0$     $p_1$   
 $\omega$  ratio:     $\omega_0 < 1$     $\omega_1 = 1$

### M2a (selection)

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

Modified from Nielsen & Yang (1998), where  $\omega_0=0$  is fixed

### M7 (beta)

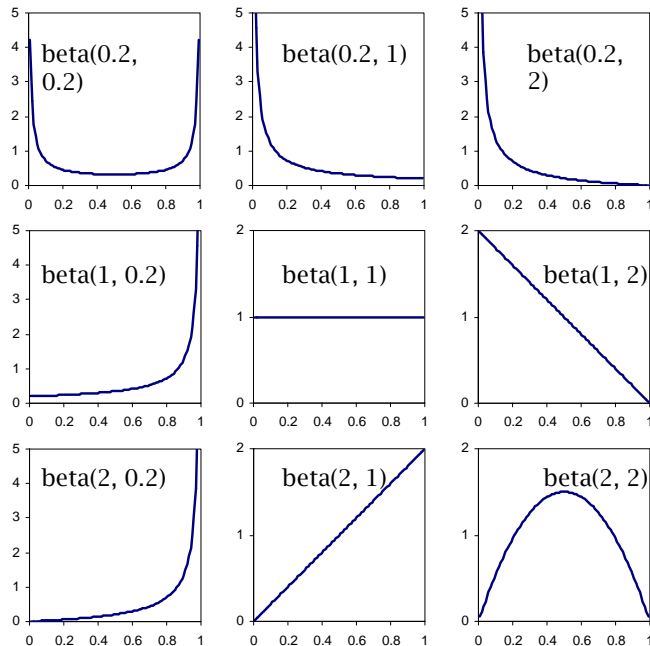
$$\omega \sim \text{beta}(p, q)$$

### M8 (beta& $\omega$ )

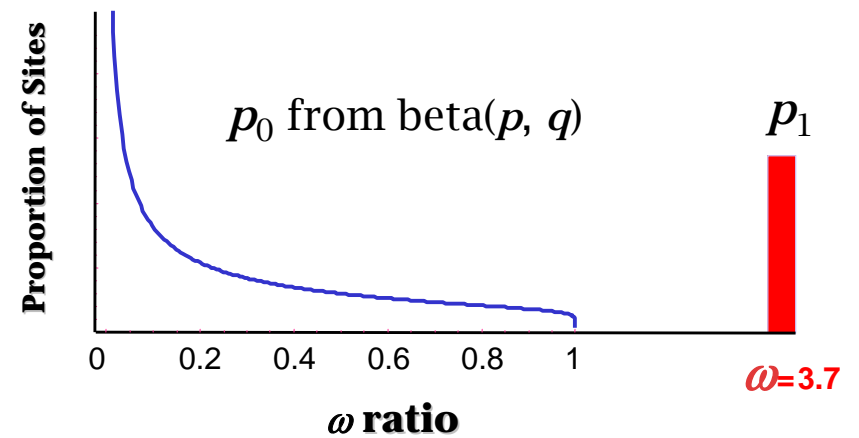
$p_0$  of sites from  $\text{beta}(p, q)$

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

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



## Mixture distribution M8(beta& $\omega$ )



## Human MHC Class I data: 192 alleles, 270 codons

Model	$\ell$	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:**  
 $2\Delta\ell = 2 \times 259.84 = 519.68, P < 0.000, \text{d.f.} = 2$

## 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)

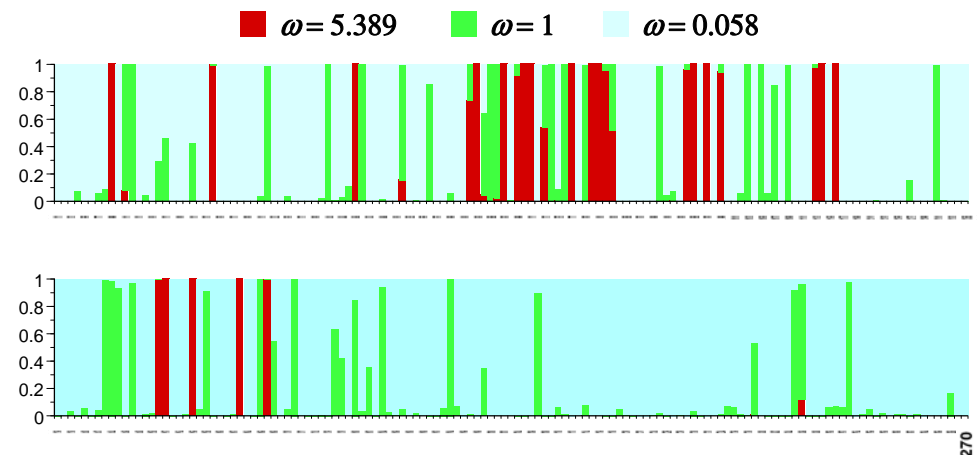
Under M2a, there are

Three site classes:  $\omega_0 = 0.058, \omega_1 = 1, \omega_2 = 5.389$

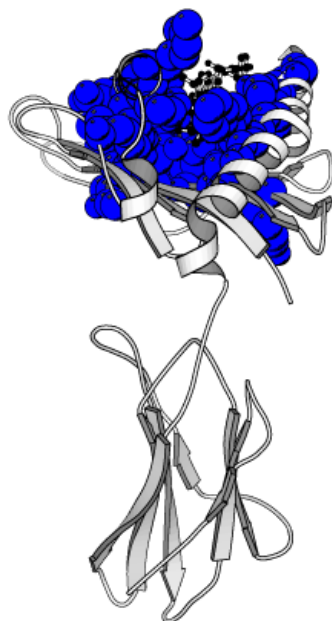
Prior proportions:  $p_0 = 0.776, p_1 = 0.140, p_2 = 0.084$

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



## Branch-site model

### Alternative hypothesis: Model A

Site class	0	1	2a	2b
Foreground	$\omega_0 < 1$	$\omega_1 = 1$	$\omega_2 \geq 1$	$\omega_2 \geq 1$
Background	$\omega_0 < 1$	$\omega_1 = 1$	$\omega_0 < 1$	$\omega_1 = 1$

### Null hypothesis: Model A with $\omega_2 = 1$

Site class	0	1	2a	2b
Foreground	$\omega_0 < 1$	$\omega_1 = 1$	$\omega_2 = 1$	$\omega_2 = 1$
Background	$\omega_0 < 1$	$\omega_1 = 1$	$\omega_0 < 1$	$\omega_1 = 1$

Yang & Nielsen. 2002. *Mol. Biol. Evol.* 19:908-917

Yang, Wong, & Nielsen. 2005. *Mol. Biol. Evol.* 22:1107-1118

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 diversification in evolution: a novel approach to protein design. *J. Biol. Chem.* 278:8733-8738.

Bielawski, J. P., K. A. Dunn, G. Sabeji, 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 TRIM5 $\alpha$  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, ).

## References, programs, etc.

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

Yang, Z., 2006. *Computational Molecular Evolution*. Oxford University Press, Oxford, England, Chapter 8

### Programs (there are also some web servers)

<http://abacus.gene.ucl.ac.uk/software/paml.html>

### Databases of positive selection

- Nickel GC, Tefft D, Adams MD. 2008. Human PAML browser: a database of positive selection on human genes using phylogenetic methods. *Nucl. Acids Res.* 36:D800–D808.
- Proux E, Studer RA, Moretti S, Robinson–Rechavi M. 2008. Selectome: a database of positive selection. *Nucl. Acids Res.*