



# Being positive about selection

---

Tõnu Margus

9. October 2006



# Objectives

---

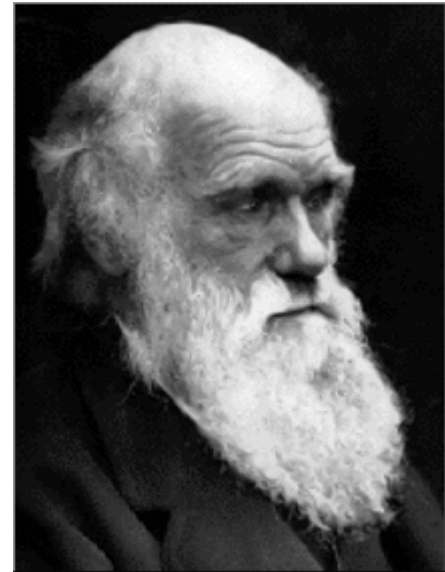
- Some members of trGTPases are aroused *via* duplication
- Universally conserved EFG has been the source of sprouting up of **tet(O,M,Q,..)**, **LepA** and **bacterial RF3**
- From my studies I have found that the EFG had a subgroup - EFG(2) and it is not possible to group it in to a single homogeneous subfamily
- That rises the question about evolutionary forces; what led to new gene families?
- The important player is **selection**, whose filter every cell's component must pass
- Inferring selection would lead us to the better understanding about the processes what are taking part after gene duplications

# Introduction

## Natural Selection

***"Natural selection is daily, hourly, scrutinising the slightest variations, rejecting those that are bad, preserving and adding up all those that are good"-***

The Origin of Species



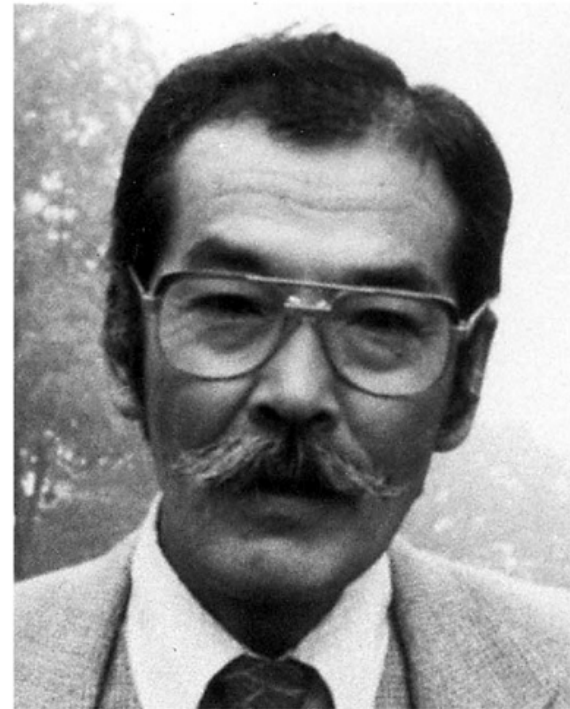
Charles Darwin

# Evolution by gene duplication

1970

***"After gene duplication  
gene copies can follow  
three possible routes:"***

- ***Nonfunctionalization***
- ***Neofunctionalization***
- ***Subfunctionalization***



*Susumu Ohno*



# Genes fate

---

- ***What ever route a gene have take, it's fate stands indissolubly bound up with selection***
- ***Insight to selection opens for us a possibility to follow and understand "how duplicated genes evolve"***
- ***Especially interesting is catch them on their way to non-, neo- or subfunctionalization***



# Selection on gene can be

---

- Negative (*Purifying*) selection
  - Housekeeping Genes'
    - Changes are bad
- Positive (*Adaptive*) selection
  - Genes that have a role in adaptation
    - Changes are good
- Genetic drift
  - Selectively neutral genes



# How to detect selection ?

---

$d_N$

Number of replacement substitutions

Number of replacement sites

$d_S$

Number of silent substitutions

Number of silent sites

$\omega = d_N/d_S > 1 \rightarrow$  Positive Selection



# Methods

---

- There are many methods for calculating  $d_N$  and  $d_S$ , they can be divided in to:
- ***Approximate methods***
  - Nei and Gojobori (1986) no tr/trv, codon bias
  - Li (1993) no codon bias
  - Ina
  - Yang and Nielsen (2000)
- ***Maximum likelihood based methods***
  - Yang



# Comparing methods

Table 2. Estimation of  $d_N$  and  $d_S$  between the human and orangutan  $\alpha_2$ -globin genes (142 codons)<sup>a</sup>

Method and/or model	$\kappa$	S	N	$d_N$	$d_S$	$d_N/d_S$ ( $\omega$ )	$f^0$	Refs
<b>Approximate methods</b>								
Nei and Gojobori (1986)	1.0	109.4	316.6	0.0095	0.0569	0.168	–	9
Li (1993)	–	NA	NA	0.0104	0.0517	0.201	–	11
Ina	2.1	119.3	299.9	0.0101	0.0523	0.193	–	14
Yang and Nielsen (2000)	6.1	61.7	367.3	0.0083	0.1065	0.078	–	15

- Ignoring tr/trv rates and codon bias leads to biased estimation of S sites and N sites
- Over- or underestimation of  $d_N$  or  $d_S$  leads to over- or underestimation of  $\omega$
- A recent ad hoc method incorporates both biases
- However, for distantly related sequences, ad hoc treatment in approximate methods can lead serious biases even under correct assumption



# A model of codon substitution

The substitution rate from codons  $i$  to  $j$  ( $i \neq j$ ) is given

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition.} \end{cases}$$

Parameter  $\kappa$  is the transition/transversion rate ratio,  $\pi_j$  is the equilibrium frequency of codon  $j$  and  $\omega$  ( $= d_N/d_S$ ) measures the selective pressure on the protein. The  $q_{ij}$  are relative rates because time and rate are confounded in such an analysis.

For example, highly biased codon usage can be caused by mutational bias and selection, and can greatly affect synonymous substitution rates. By employing parameters  $\pi_j$  for the frequency of codon  $j$  in the model estimation of substitution rates will fully account for codon-usage bias, irrespective of its source.



# Maximum Likelihood methods

---

- These methods based on explicit models of codon substitutions
- Parameters (seq. divergence,  $\omega$ ) for these models are estimated from data by ML
- Model is formulated at instantaneous rates (multiple changes are not allowed)
- Probability theory accomplishes all difficult tasks in one step; estimating parameters, correcting for multiple hits, weighting pathways for changing codon's
- Statistical tests can be used to test whether  $d_N$  is significantly higher than  $d_S$
- A likelihood test can be used for testing two alternative hypothesis
  - null model where  $\omega$  is fixed and
  - more complicated model where  $\omega$  is set as free parameter

# Comparing methods

These seq are 10% different at silent sites and 1 % at nonsynonymous sites

**Table 2. Estimation of  $d_N$  and  $d_S$  between the human and orangutan  $\alpha_2$ -globin genes (142 codons)<sup>a</sup>**

Method and/or model	$\kappa$	S	N	$d_N$	$d_S$	$d_N/d_S$ ( $\omega$ )	$\ell^o$	Refs
<b>Approximate methods</b>								
Nei and Gojobori (1986)	1.0	109.4	316.6	0.0095	0.0569	0.168	–	9
Li (1993)	–	NA	NA	0.0104	0.0517	0.201	–	11
Ina	2.1	119.3	299.9	0.0101	0.0523	0.193	–	14
Yang and Nielsen (2000)	6.1	61.7	367.3	0.0083	0.1065	0.078	–	15
<b>ML methods<sup>b</sup></b>								
(1) Fequal, $\kappa = 1$	1.0	108.5	317.5	0.0093	0.0557	0.167	–633.67	16
(2) Fequal, $\kappa$ estimated	3.0	124.6	301.4	0.0099	0.0480	0.206	–632.47	16
(3) F1×4, $\kappa = 1$ fixed	1.0	129.1	296.9	0.0092	0.0671	0.137	–612.40	16
(4) F1×4, $\kappa$ estimated	3.9	137.1	288.9	0.0093	0.0624	0.149	–610.48	16
(5) F3×4, $\kappa = 1$ fixed	1.0	63.2	362.8	0.0084	0.0973	0.087	–560.76	16
(6) F3×4, $\kappa$ estimated	5.4	60.6	365.4	0.0084	0.1061	0.079	–557.85	16
(7) F61, $\kappa = 1$ fixed	1.0	58.3	367.7	0.0082	0.1145	0.072	–501.39	16
(8) F61, $\kappa$ estimated	5.3	55.3	370.7	0.0082	0.1237	0.066	–498.61	16

However, for distantly related sequences, *ad hoc* treatment in approximate methods can lead to serious biases even under the correct assumptions

<sup>o</sup> $\ell$  is the log-likelihood value.



# Summary of first part

---

- First challenge for reliable method is to separate selection on nucleic acid/gene level from selection on protein level
- Assumptions have a greater effect than used method
- The pair wise comparison has few power because it averages the  $\omega$  ratio over sites and over time

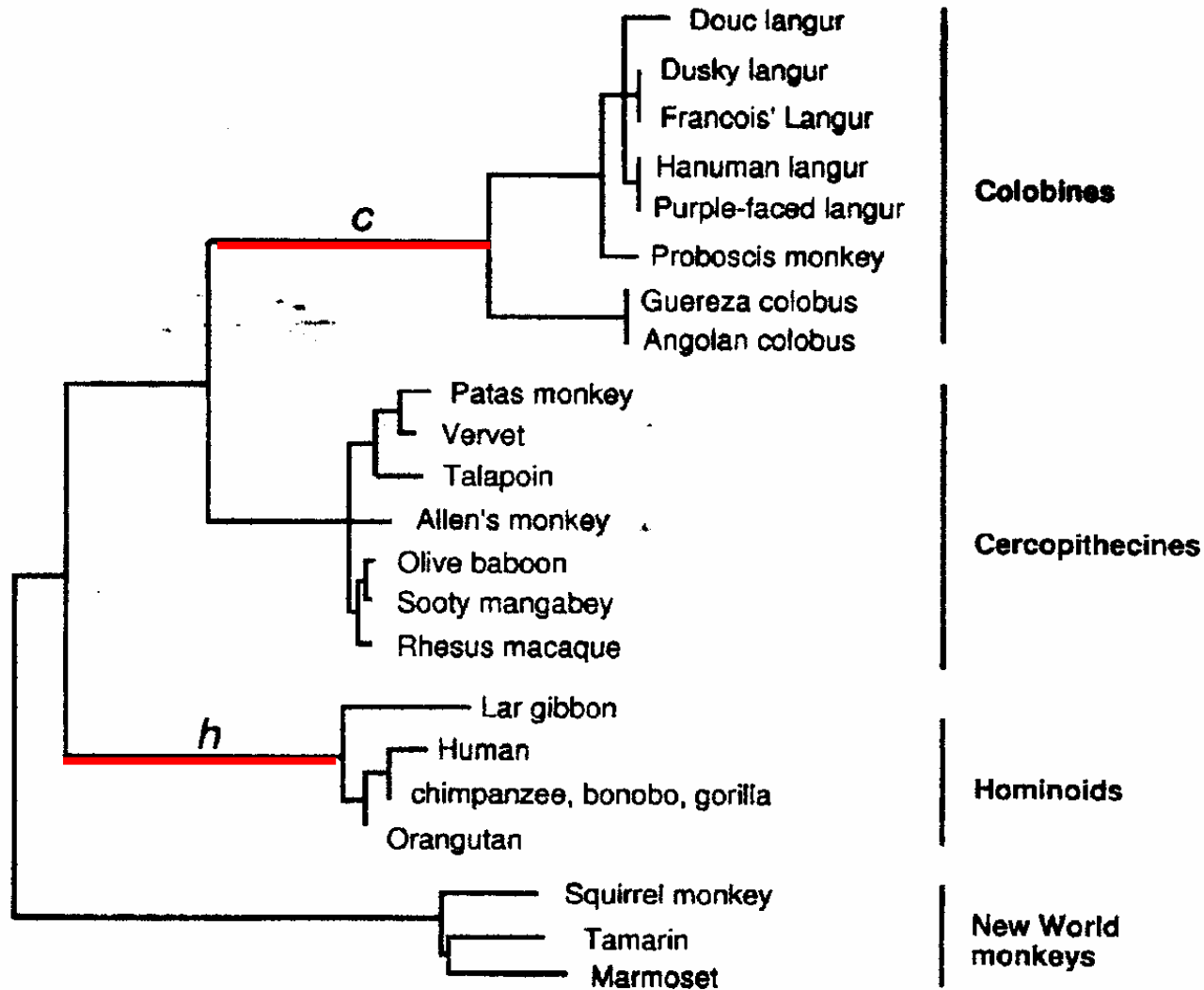


# Detecting selection in more detail

---

- Detecting lineage-specific episodes of **positive** Darwinian **selection**
  - If some species have moved to new environment, positive Darwinian selection indicates an adaptation with a new environment
- Detecting sites evolving under **positive selection**
  - Variability on certain sites is advantageous
  - For example on parasite host relationship helps parasite to avoid hosts immune response
- Detecting **functional divergence** at individual codon sites
  - Combines site specific and lineage specific approach
  - Address the question rather about changing selective pressure to specific position than about detecting positive selection

# Heterogeneous model of evolution



Based on the given phylogeny, and from the previously known results (Messier and Stewart), we can formulate the hypotheses that can be tested using maximum likelihood.



$\omega_0$  is the background  $d_N/d_S$  ratio

$\omega_C$  is the  $d_N/d_S$  ratio of Branch C (colobines)

$\omega_H$  is the  $d_N/d_S$  ratio of Branch H (human)

### **Then test the interesting hypotheses:**

Every  $\omega_j$  is different:

$$\omega_0 = \omega_C = \omega_H$$

$$\omega_0 = \omega_C, \omega_H$$

$$\omega_0, \omega_C, \omega_H$$

etc....

“free-ratio” model

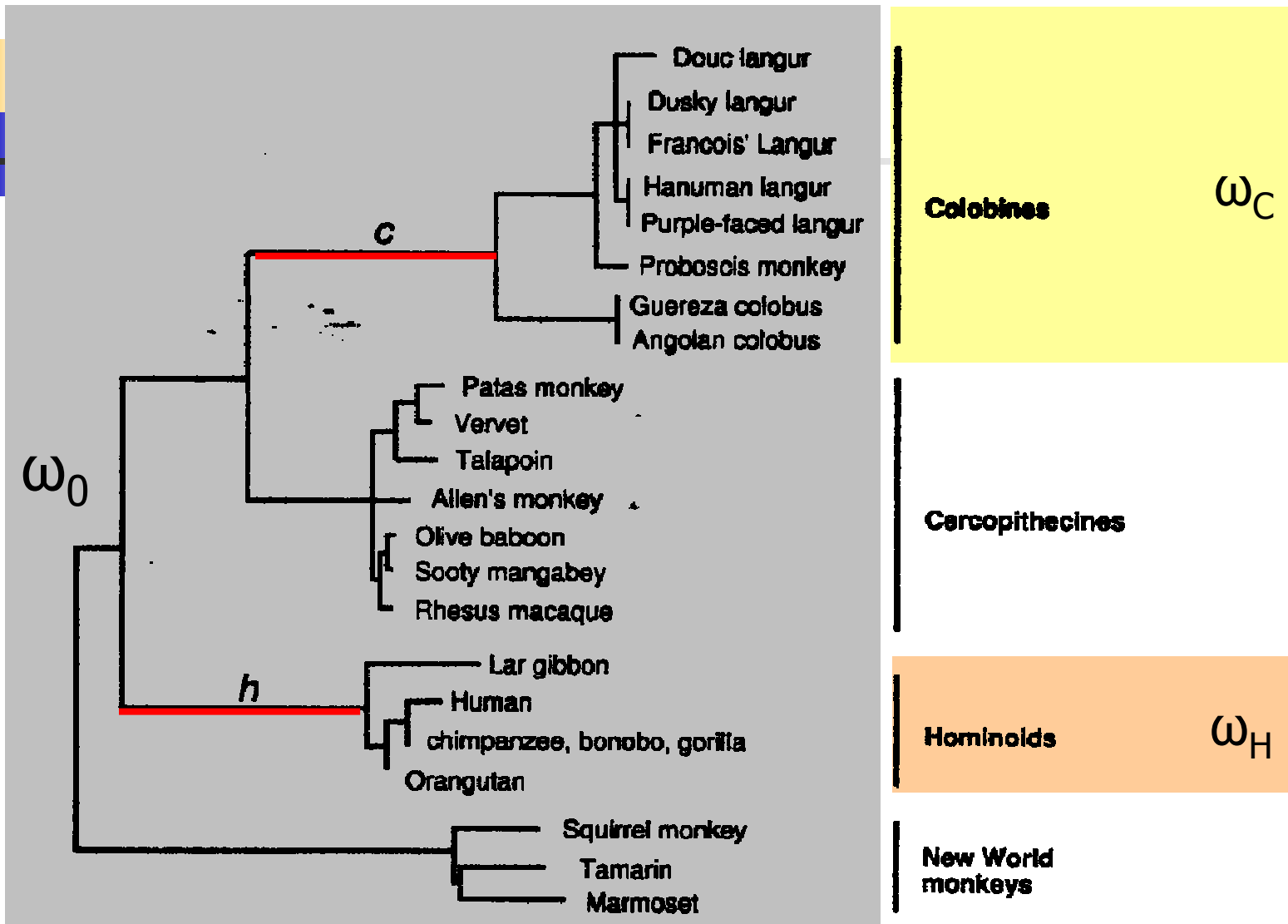
“one-ratio” model

“two-ratio” model

“three-ratio” model



# Heterogeneous model of evolution



# Results from Ziheng Yang 1998

- 1) The  $d_N/d_S$  ratios in the Primate Lysozyme genes are highly variable among evolutionary lineages, indicating that the evolution of primate Lysozyme is incompatible with a neutral model
- 2) The  $d_N/d_S$  of the lineage leading to the Hominids was significantly greater than 1
- 3) The  $d_N/d_S$  leading to the colobines was significantly greater than the background  $d_N/d_S$  ratio, but was not greater than 1
- 4) Methods for detecting selection along lineages work only if the  $\omega$  ratio averaged over all sites is  $>1$



# Sites under positive selection

---

- Previous approaches effectively averaged  $\omega$  ratio across all sites
- Positive selection detected, when this average is  $> 1$  (it is conservative test)
- Variable regions in DNA are not always junk regions
- If the variability have driven by positive selection, the functional importance have proved



# Approaches for detecting sites under positive selection

---

- Fitch *et al.* (1997)
- Suzuki and Gojobori (1999)

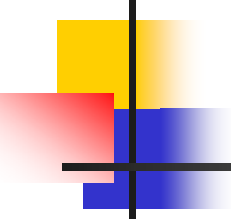
Both methods are using reconstruction of ancestral state and using it as real data, what is most unreliable at positions under positive selection!

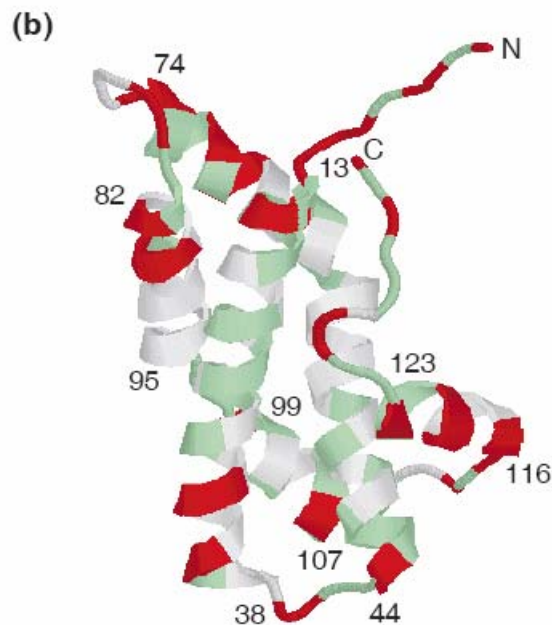
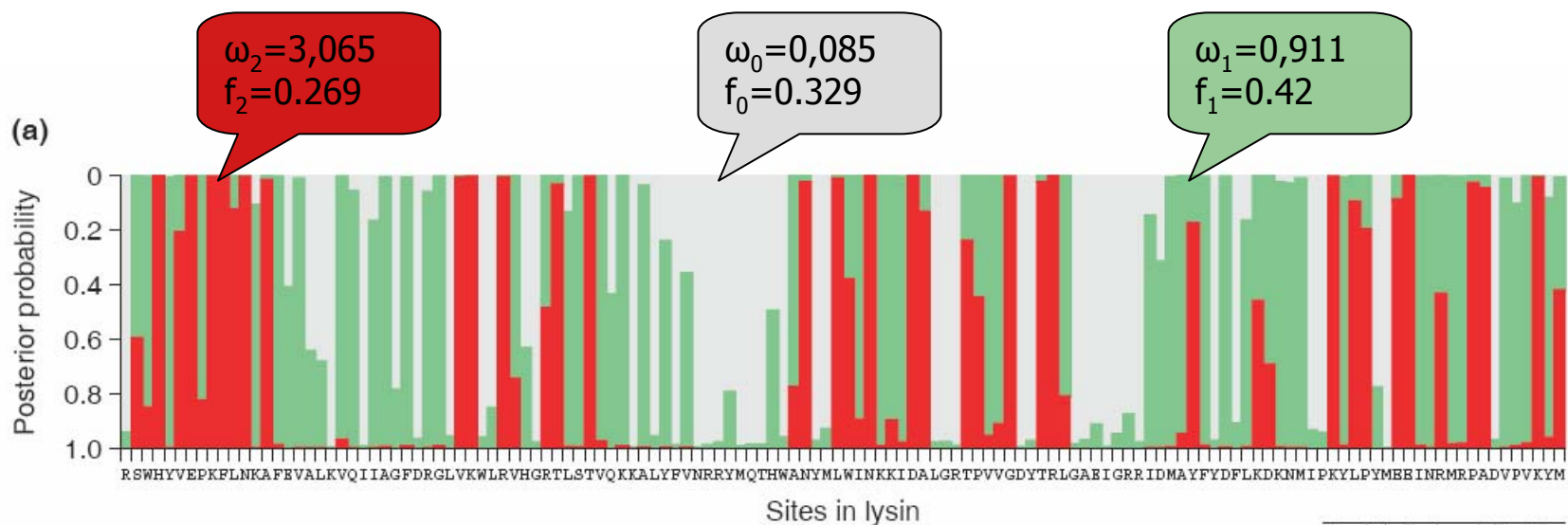


# On likelihood based methods

---

- The standard approach is to use a statistical distribution to describe the variation of  $\omega$  among sites
- The test of positive selection then involves two major steps:
  - to test whether sites exist where  $\omega > 1$ , which is achieved by a likelihood-ratio test comparing a model that does not allow for such sites with a more general model that does
  - to use the Bayes theorem to identify positively selected sites when they exist

- 
- 
- The null model, **M1** (neutral), assumes a class of conserved sites with  $\omega=0$  and another class of neutral sites with  $\omega=1$
  - The alternative model, **M2** (selection), adds a third class of sites with  $\omega$  estimated from the data.
  - If **M2** fits the data significantly better than **M1** and the estimated  $\omega$  ratio for the third class in **M2** is  $>1$ , then some sites are under **positive selection**
  - Zanutto *et al.* used this test to identify several sites under strong positive selection in the *nef* gene of HIV, whereas both pairwise comparison and sliding window analysis failed



**Fig. 1.** The identification of sites under positive selection from the sperm lysin genes of 25 abalone species. (a) Posterior probabilities for site classes with different  $\omega$  ratios along the sequence. The lysin sequence of the red abalone (*Haliotis rufescens*) is shown below the x-axis. ML estimates under Model M3 (discrete)<sup>37</sup> suggest three site classes with the  $\omega$  ratios at  $\omega_0 = 0.085$  (grey),  $\omega_1 = 0.911$  (green) and  $\omega_2 = 3.065$  (red), and with proportions  $p_0 = 0.329$ ,  $p_1 = 0.402$  and  $p_2 = 0.269$ . These proportions are the prior probabilities (Box 1) that any site belongs to the three classes. The data (codon configurations in different species) at a site alter the prior probabilities dramatically, and thus the posterior probabilities might be different from the prior probabilities. For example, the posterior probabilities for Site 1 are 0.944, 0.056 and 0.000, and thus this site is likely to be under strong purifying selection. The posterior probabilities for Site 4 are 0.000, 0.000 and 1.000, and thus this site is almost certainly under diversifying selection. (b) Lysin crystal structure from the red abalone (Protein Data Bank file 1LIS), with sites coloured according to their most likely class inferred in (a). The structure starts from amino acid four (His) at the N-terminus, because the first three amino acids are unresolved. The five  $\alpha$ -helices are indicated:  $\alpha_1$  from amino acids 13 to 38,  $\alpha_2$  from 44 to 74,  $\alpha_3$  from 82 to 95,  $\alpha_4$  from 99 to 107 and  $\alpha_5$  from 116 to 123. Note that sites potentially under positive selection (red) are scattered all over the primary sequence but tend to cluster around the top and bottom of the crystal structure. *Reproduced, with permission, from Ref. 39.*



# Limitations of current methods

---

- Methods for detecting positive selection at sites works only if the  $\omega$  ratio averaged over all branches **is >1**
- Constancy of selective pressure at sites appears to be a much more serious assumption than constancy among lineages, especially for genes likely to be under continuous selective pressure, such as the HIV *env* gene
- Models that allow  $\omega$  to vary among both lineages and sites should have increased power





# Detecting functional divergence at individual codon sites

Bielawski JP. and Z. Yang JME 2004

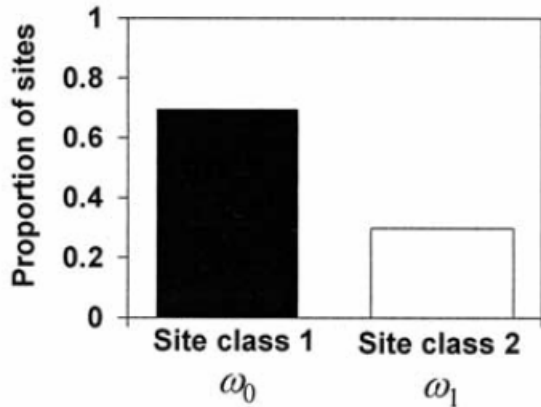
## **We assume that:**

- selective pressure varies among the amino acid sites
- a subset of a sites experience a change in selective pressure at a point in evolutionary history, such as a duplication event

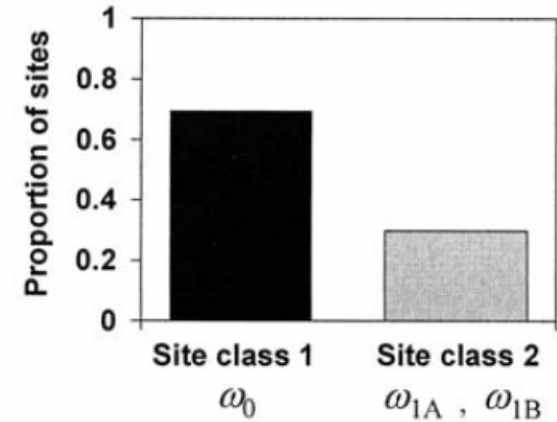
## **We:**

- don't know the history of selective pressure
- wish to identify which sites have experienced a change following the duplication

**A. Discrete model (M3) with  $k = 2$**



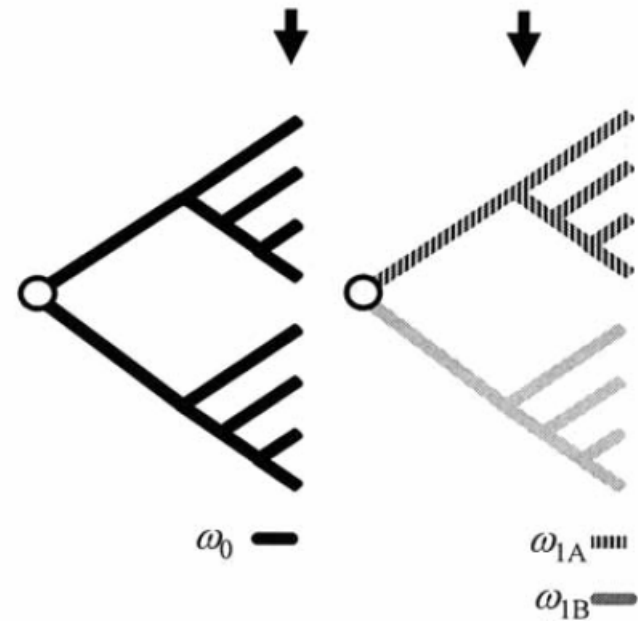
**B. Model D with  $k = 2$**

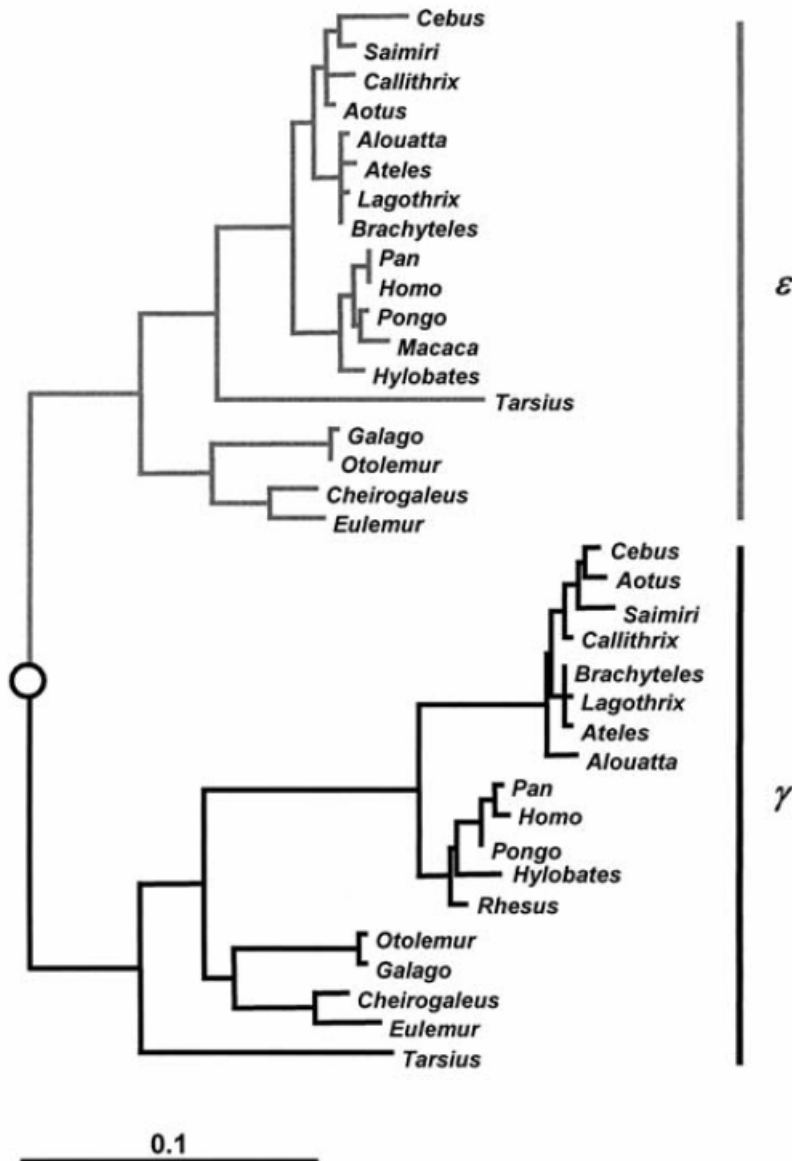


**M3 (discrete)**

No of parameters  $2k-1;$   
 $k$  ( $k=3$ )  
 Parameters  $f_0, f_1, \dots, f_{(k-1)}$   
 $\omega_0, \omega_1, \dots, \omega_{(k-1)}$

Model **D** extends model M3 by allowing selective pressure at one class of sites to differ in different parts of a phylogeny





$\epsilon$  and  $\gamma$  arose about 80–100 MYA via a tandem duplication of an embryonic  $\epsilon$ -type globin (Koop and Goodman 1988).

Expression of  $\epsilon$  is embryonic in all placental mammals, while  $\gamma$  expression is embryonic only in nonprimate placental mammals and prosimian primates (Johnson et al. 1996).

Persistence of both  $\epsilon$  and  $\gamma$  over 80 to 100 million years of evolution implies strong selective pressure for both gene products, presumably due to some form of genetic co-option and divergence.

If functions of  $\epsilon$  and  $\gamma$  had not diverged, it is likely that one copy would have become nonfunctional

### The objective of this analysis was:

- test for divergence in selective pressure between  $\epsilon$  and  $\gamma$
- identify sites consistent with this type of selective pressure if they existed.

Fig. 2. Gene tree for the 36 sequences from the  $\epsilon$  and  $\gamma$  gene family. The topology was obtained by using maximum likelihood



# Results

---

- The one ratio model (M0) yielded an estimated  $\omega=0.19$ , indicating that purifying selection is dominated the evolution of these globins
- An LRT among M0 and M3 indicating significant variation in selective pressure among sites (M3 with  $k=2$ )
- An LRT of M3 with  $k=2$  against M3 with  $k=3$  site classes was not significant.
- M3 with  $k=2$  suggests a large fraction of sites (70%) evolving under strong purifying selection ( $\omega=0.05$ ) and a smaller fraction of sites (30%) evolving more quickly ( $\omega=0.55$ ).

# Testing divergence in selective pressure between $\epsilon$ and $\gamma$ globins

Applying the new **model D** which accommodates both the **heterogeneity among the sites** and **divergent selective pressure**

- Significance of the LRT for divergent selective pressure for was borderline ( $k=2$ ,  $P=0.05$ ) and unmistakable when  $k=3$ , ( $P=0.001$ )
- Model D with  $k=3$  suggests following proportion of sites:
  - $\sim 65\%$   $\omega=0,04$
  - $\sim 19\%$   $\omega=0,61$
  - $\sim 16\%$   $\omega_{2\epsilon}=0,008$  and  $\omega_{2\gamma}=0,79$

12 sites with posterior probabilities  $\geq 75\%$

On 3D structure,

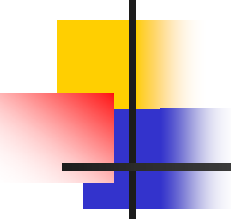
- 4 amino acids mapped at or within tetramer interface
- 2 are located at binding site of DPG
- 4 were located at heme binding site



# Conclusions

---

- ML methods successfully subtract selection on nucleic acid level from selection on protein level
- Statistical tests is used to test whether  $d_N$  is significantly higher than  $d_S$
- A likelihood test is used for testing two alternative hypothesis
- There is a rich set of models for testing different evolutionary scenarios



---

Thank you for your  
attention!