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

Definitions

\[d_S(K_s)\]: number of synonymous substitutions per synonymous site

\[d_N(K_a)\]: number of nonsynonymous substitutions per nonsynonymous site

\[\omega = \frac{d_N}{d_S}\]: nonsynonymous/synonymous rate ratio
**Counting method**

1. If we expect $M:S$ to be 74.5:25.5% before selection on the protein, and observe 5:5 substitutions (differences), then
   \[ \omega = \frac{d_N}{d_S} = \frac{5/5}{74.5/25.5} = 0.34 \]

2. The gene is 3×300 nucleotides long, so
   \[ S = 900 \times 25.5\% = 229.5 \]
   \[ N = 900 \times 74.5\% = 670.5 \]
   \[ d_S = \frac{5}{229.5} = 0.0218 \]
   \[ d_N = \frac{5}{670.5} = 0.0075 \]

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

- 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

**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:

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Syn</th>
<th>Nonsyn</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCT (Pro) ↔ CAT (His) ↔ CAG (Gln)</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>CCT (Pro) ↔ CCG (Pro) ↔ CAG (Gln)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Average</td>
<td>0.5</td>
<td>1.5</td>
</tr>
</tbody>
</table>

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)

\[ d_N = d_s \times \omega \]

Yang (2006 *Computational Molecular Evolution*, Chapter 2)
**Counting method**


**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**
- $\text{CTC (Leu)} \rightarrow \text{CTG (Leu)}$: $\pi_{CTG}$
- $\text{TTG (Leu)} \rightarrow \text{CTG (Leu)}$: $\kappa \pi_{CTG}$

**Nonsynonymous**
- $\text{CTG (Val)} \rightarrow \text{CTG (Leu)}$: $\omega \pi_{CTG}$
- $\text{CCG (Pro)} \rightarrow \text{CTG (Leu)}$: $\kappa \omega \pi_{CTG}$

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

<table>
<thead>
<tr>
<th>Method/Model</th>
<th>$\kappa$</th>
<th>$S$</th>
<th>$N$</th>
<th>$d_N$</th>
<th>$d_S$</th>
<th>$d_S/d_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NG86</td>
<td>1</td>
<td>109.4</td>
<td>316.6</td>
<td>0.0095</td>
<td>0.0569</td>
<td>0.168</td>
</tr>
<tr>
<td>Ina95</td>
<td>2.1</td>
<td>119.3</td>
<td>299.9</td>
<td>0.0101</td>
<td>0.0523</td>
<td>0.193</td>
</tr>
<tr>
<td>YN00</td>
<td>6.1</td>
<td>61.7</td>
<td>367.3</td>
<td>0.0083</td>
<td>0.1065</td>
<td>0.078</td>
</tr>
</tbody>
</table>

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

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.

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

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

$$\hat{p} = \frac{x}{n} = 0.1$$

ML estimation of $d_S$ and $d_N$

The probability of observing a site with codons $i$ and $j$ in the two sequences is $\pi_i \rho_{ij}(t)$

The log likelihood is

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

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 $N$) are calculated from $q_{ij}$ by fixing $\omega = 1$.

Chapman-Kolmogorov theorem

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

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

$$\Pr(ij \mid t_1, t_2) = \sum_k \pi_k p_{ki}(t_1)p_{kj}(t_2)$$

which is the same requirement as

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

← Chapman-Kolmogorov theorem
### Human and orangutan $\alpha_2$-globin genes (142 codons)

<table>
<thead>
<tr>
<th>Method/Model</th>
<th>$\kappa$</th>
<th>$S$</th>
<th>$N$</th>
<th>$d_S$</th>
<th>$d_S$</th>
<th>$d_S/d_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NG86</td>
<td>1</td>
<td>109.4</td>
<td>316.6</td>
<td>0.0095</td>
<td>0.0569</td>
<td>0.168</td>
</tr>
<tr>
<td>Ina95</td>
<td>2.1</td>
<td>119.3</td>
<td>299.9</td>
<td>0.0101</td>
<td>0.0523</td>
<td>0.193</td>
</tr>
<tr>
<td>YN00</td>
<td>6.1</td>
<td>61.7</td>
<td>367.3</td>
<td>0.0083</td>
<td>0.1065</td>
<td>0.078</td>
</tr>
</tbody>
</table>

ML (GY94)
(1) ML Fequal, $\kappa = 1$
$1 \quad 108.5 \quad 317.5 \quad 0.0093 \quad 0.0557 \quad 0.167$
(2) ML Fequal, $\kappa$ estimated
$3.0 \quad 124.6 \quad 301.4 \quad 0.0099 \quad 0.0480 \quad 0.206$
(7) ML F61, $\kappa = 1$ fixed
$1 \quad 58.3 \quad 367.7 \quad 0.0082 \quad 0.1145 \quad 0.072$
(8) ML F61, $\kappa$ estimated
$5.3 \quad 55.3 \quad 370.7 \quad 0.0082 \quad 0.1237 \quad 0.066$

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Counting methods</strong></td>
<td>MEGA; codeml &amp; yn00 in PAML</td>
</tr>
<tr>
<td>NG86</td>
<td>MEGA, DAMBE, codeml</td>
</tr>
<tr>
<td>Li93</td>
<td>DIVERGE by Comeron</td>
</tr>
<tr>
<td>Comeron 95</td>
<td>yn00 in PAML</td>
</tr>
<tr>
<td>YN00</td>
<td></td>
</tr>
<tr>
<td><strong>ML methods</strong></td>
<td>codeml</td>
</tr>
<tr>
<td>GY94</td>
<td></td>
</tr>
</tbody>
</table>
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)


Positive & negative selection

<table>
<thead>
<tr>
<th>Genotype</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>$p^2$</td>
<td>$2p(1-p)$</td>
<td>$(1-p)^2$</td>
</tr>
<tr>
<td>Fitness</td>
<td>1</td>
<td>$1+s$</td>
<td>$1+2s$</td>
</tr>
</tbody>
</table>

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

$s$ is selection coefficient:

$s = 0$: neutral evolution
$s < 0$: negative (purifying) selection
$s > 0$: positive selection (adaptive evolution)
Theories of molecular evolution

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.

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.

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)

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

$$\omega = \frac{d_N}{d_S} = \frac{5}{74.5\%/25.5\%} = 0.34$$
Codon–substitution model ($Q_{61 \times 61}$): Rates to CTG

**Synonymous**
- CTC (Leu) → CTG (Leu)
- TCTG (Leu) → CTG (Leu)

**Nonsynonymous**
- GTG (Val) → CTG (Leu)
- CCG (Pro) → CTG (Leu)

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

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

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

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

**Codon substitution models**


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

Likelihood ratio test to compare two nested models

Adaptive evolution in primate lysozyme

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.

Log-likelihood values and parameter estimates

<table>
<thead>
<tr>
<th>Model</th>
<th>p</th>
<th>$\ell$</th>
<th>$\omega_0$</th>
<th>$\omega_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. 1-ratio: $\omega_0 = \omega_C$</td>
<td>35</td>
<td>-1043.84</td>
<td>0.574</td>
<td>$\omega_0$</td>
</tr>
<tr>
<td>B. 2-ratios: $\omega_0$, $\omega_C$</td>
<td>36</td>
<td>-1041.70</td>
<td>0.489</td>
<td>3.383</td>
</tr>
<tr>
<td>C. 2-ratios: $\omega_0$, $\omega_C = 1$</td>
<td>35</td>
<td>-1042.50</td>
<td>0.488</td>
<td>1</td>
</tr>
</tbody>
</table>

Data from Messier & Stewart 1997 Nature 385: 151-154)

Likelihood ratio test statistics

<table>
<thead>
<tr>
<th>Null hypothesis</th>
<th>$2\Delta\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega_C = \omega_0$</td>
<td>4.24*</td>
</tr>
<tr>
<td>$\omega_C = 1$</td>
<td>1.60</td>
</tr>
</tbody>
</table>

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

- Focus on sites potentially under selection based on structure

- Use a statistical distribution to model the $\omega$ variation
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

Two pairs of useful models

M1a (neutral)
Site class: 0  1
Proportion: $\rho_0 \quad \rho_1$
$\omega$ ratio: $\omega_0 < 1 \quad \omega_1 = 1$

M2a (selection)
Site class: 0  1  2
Proportion: $\rho_0 \quad \rho_1 \quad \rho_2$
$\omega$ ratio: $\omega_0 < 1 \quad \omega_1 = 1 \quad \omega_2 > 1$

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

M7 (beta)
$\omega \sim \text{beta}(\rho, \varphi)$

M8 (beta&$\omega$)
$\rho_0$ of sites from $\text{beta}(\rho, \varphi)$
$\rho_1 = 1 - \rho_0$ of sites with $\omega > 1$

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

Mixture distribution

$\rho_0$ from $\text{beta}(\rho, \varphi)$
$\rho_1$

$\omega$ ratio

$\omega=3.7$
Human MHC Class I data:
192 alleles, 270 codons

<table>
<thead>
<tr>
<th>Model</th>
<th>$\ell$</th>
<th>Parameter estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1a (neutral)</td>
<td>−7,490.99</td>
<td>$p_0 = 0.830$, $\omega_0 = 0.041$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$p_1 = 0.170$, $\omega_1 = 1$</td>
</tr>
<tr>
<td>M2a (selection)</td>
<td>−7,231.15</td>
<td>$p_0 = 0.776$, $\omega_0 = 0.058$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$p_1 = 0.140$, $\omega_1 = 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$p_2 = 0.084$, $\omega_2 = 5.389$</td>
</tr>
</tbody>
</table>

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.


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

<table>
<thead>
<tr>
<th>Site class</th>
<th>0</th>
<th>1</th>
<th>2a</th>
<th>2b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreground</td>
<td>$\omega_0 &lt; 1$</td>
<td>$\omega_1 = 1$</td>
<td>$\omega_2 \geq 1$</td>
<td>$\omega_2 \geq 1$</td>
</tr>
<tr>
<td>Background</td>
<td>$\omega_0 &lt; 1$</td>
<td>$\omega_1 = 1$</td>
<td>$\omega_0 &lt; 1$</td>
<td>$\omega_1 = 1$</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Site class</th>
<th>0</th>
<th>1</th>
<th>2a</th>
<th>2b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreground</td>
<td>$\omega_0 &lt; 1$</td>
<td>$\omega_1 = 1$</td>
<td>$\omega_2 = 1$</td>
<td>$\omega_2 = 1$</td>
</tr>
<tr>
<td>Background</td>
<td>$\omega_0 &lt; 1$</td>
<td>$\omega_1 = 1$</td>
<td>$\omega_0 &lt; 1$</td>
<td>$\omega_1 = 1$</td>
</tr>
</tbody>
</table>


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.


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.


Programs (there are also some web servers)

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

Databases of positive selection