Today’s lecture:
The evolutionary detective - genome duplication & detecting the ‘invisible hand’ of natural selection

The duplication story: *deja vu* all over again

- The Menu Bar:
- Part 1: Evolution and duplication - basic background: why is detecting selection hard?
- Part 2: Detecting selection
  - Simple counting methods & their corrections
  - Codon substitution methods
  - Maximum Likelihood methods
- Part 3: Applications
Duplication is everywhere: Whole genome duplication in vertebrates (Nature, Oct 2004)

Arabidopsis possible genome duplication

- Segmentally duplicated regions in the Arabidopsis genome.
- Much of the genome is in pairs (not in triplicates, etc.).
Human contains up to 4 copies of many Drosophila genes.
Internal Gene Duplication

- **Ancestral Trypsinogen gene**
- **Antifreeze glycoprotein gene**

Terminology

We take homologyTMD (here of proteins) to mean the explanation of an observed similarity in virtue of common ancestry.

A common mode of protein evolution is by duplication. Depending on the relations between duplication and speciation dates, we have two different types of homologous proteins:

- **Orthologs**: the “same” gene in different organisms; common ancestry goes back to a speciation event.
- **Paralogs**: different genes in the same organism; common ancestry goes back to a gene duplication.

Lateral gene transfer gives another possibility for observed similarity accounted for by common ancestry.

homologyTMD = copyright jahouse
“Paralog’ - genes whose similarity can be explained in virtue of a common, earlier duplication event

Paralogs in gene families

Glutamate-1-semialdehyde aminotransferase (COG0001)

Data taken from the COGs database: www.ncbi.nlm.nih.gov/COG

Paralogs are everywhere!

The first 126 COGs (Clusters of Orthologous Groups). Each row represents a gene family.
Part 1: Methodological preliminaries

- Part 1a: The forces of evolution: mutation, selection, migration
- The signal to noise problem: Why is measuring the effect of selection hard?
- Part 1b: Back to first grade - How do we measure selection? - counting

Two views of evolution

“I have called this principle, by which each slight variation, if useful, is preserved, by the term Natural Selection.” C. Darwin, Origin of Species, 1859

Evolution by natural selection
Ohno's answer

"natural selection merely modified, while redundancy created"

The consequences of gene duplications

When combined with geographical isolation, these small-scale gene rearrangements may contribute to the emergence of new reproductively isolated species.


Duplication of genes at unlinked chromosomal locations can passively give rise to a small-scale chromosomal rearrangement.

Kimura and Ohta’s Laws of Molecular Evolution

1. For each protein, the rate of evolution in terms of amino acid substitutions is approximately constant per year per site for various lines, as long as the function and tertiary structure of the molecule remain essentially unaltered.
2. Functionally less important molecules or parts of molecules evolve (in terms of mutant substitutions) faster than more important ones.
3. Those mutant substitutions that are less disruptive to the existing structure and function of the molecule (conservative substitutions) occur more frequently in evolution than more disruptive ones.
4. Gene duplication must always precede the emergence of a gene having a new function.
5. Selective elimination of definitely deleterious mutant and random fixation of selectively neutral or very slightly deleterious mutants occur far more frequently in evolution than positive Darwinian selection of definitely advantageous mutants.

Kimura and Ohta (1974) PNAS 71: 2848-2852
Terminology
We take homology\textsuperscript{TM} (here of proteins) to mean the explanation of an observed similarity in virtue of common ancestry.

A common mode of protein evolution is by duplication. Depending on the relations between duplication and speciation dates, we have two different types of homologous proteins:

- **Orthologs**: the “same” gene in different organisms; common ancestry goes back to a speciation event.
- **Paralogs**: different genes in the same organism; common ancestry goes back to a gene duplication.

Lateral gene transfer gives another possibility for observed similarity accounted for by common ancestry.

The challenge

**synonymous**: TTT (Phe) $\rightarrow$ TTC (Phe)

**nonsynonymous**: TTT (Phe) $\rightarrow$ TTA (Leu)

Q: Why???
A: Signal to noise problem
A: dynamical system problem

These are the facts of life!
Selection is a weak signal amidst much noise!

Signal: force of selection, $s$, measured by fluctuations in nucleotide frequencies

But there are several other equally compelling (and stronger) forces that can change nt frequencies: migration, $m$, mutation, $u$, and drift (vs. pop size $N$)

The upshot from population genetics theory is that we can compare these forces in terms of the following, which indicate when a particular force is strong enough to ‘drive’ a gene/population to fixation:

- $Ns > 1$ (size of $s$? approx $10^{-3}$)
- $Nm > 1$ (size of $m$? approx $1$/generation)
- $Nu > 1$ (size of $u$? approx $10x$ as $s$, per gene)

Signal-to-noise

Problem 1: the signal, the force of selection, $s$, measured by fluctuations in nucleotide frequencies is weak compared to the other forces

Problem 1b: add to this the background ‘white noise’ of constant ‘neutral’ mutations - the churn

Problem 2: the problem of recovering the signal $s$ from the noise is ill-posed in a dynamical system sense

We cannot observe $s$ in operation - too slow

To recover $s$ as a force pushing the trajectory of nt changes over time, we need to know: initial state, final state.

We can observe only the final state
The solution

Problem 1: must dig hard
Problem 1a: don’t expect too much; must pick and choose your battles
Problem 2: assume stochastic steady state (at equilibrium at moment of observation)

Part 1b: back to elementary school - counting as the foundation for measuring selection

- At bottom, all evolution occurs by nucleotide (nt) changes
- So ‘all’ we have to do is to compare two sequences and count differences...
- But...how do we compare sequences? what do we count? Nucleotides? codons? amino acids?
- Do all differences count equally? Do all sites?
- Because: nt changes → amino acid changes→ protein changes → phenotype (what is selected)
Basic counting - let’s start with nucleotides
Idea: # of evolutionary events \( \equiv \) # of differences

**Part 1: Complications**

**Step 2: Counting differences**

<table>
<thead>
<tr>
<th>human</th>
<th>CCT</th>
<th>TCT</th>
<th>CCT</th>
<th>GCC</th>
<th>GAC</th>
<th>AAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cow</td>
<td>CAG</td>
<td>...</td>
<td>G.C</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Rabbit</td>
<td>...</td>
<td>T</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Rat</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>A</td>
<td>AT</td>
<td>A</td>
</tr>
<tr>
<td>Mouse</td>
<td>...</td>
<td>C</td>
<td>G.G</td>
<td>...</td>
<td>T</td>
<td>...</td>
</tr>
</tbody>
</table>

But this idea is too simple...! Why?

Redundancy in the table creates synonymous vs. nonsynonymous possible changes
Some codon changes are AA ‘silent’ (Pauling; Zuckerkandl, 1959 (!!))

synonymous: TTT (Phe) → TTC (Phe)

nonsynonymous: TTT (Phe) → TTA (Leu)
Nucleotide differences come in two flavors: Synonymous and Nonsynonymous

**Synonymous nt change** (dS, K_S) = change in nt does *not* change amino acid, hence protein & function *stays the same* (neutral wrt selection)

**Nonsynonymous nt change** (dN, K_A) = change in nt *does* change the amino acid, hence protein changes (and so function?)

So, we must adjust for this: compare dN vs. dS (or K_A vs. K_S) to get a ratio test for selection:

### Selection Definitions

dS = rate of synonymous substitution
dN = rate of nonsynonymous substitution

\[
dN/dS < 1 \text{ negative selection}
\]

\[
dN/dS = 1 \text{ neutral evolution}
\]

\[
dN/dS > 1 \text{ positive selection}
\]
dN = # nonsynonymous substitutions/# nonsynonymous sites

dS = # synonymous substitutions/# synonymous sites

The ratio dN/dS, \( \omega \), measures the selective pressure.

---

Test for selection by comparing dN and dS

- \( \omega = 1 \): Neutral selection (background neutral)
- \( \omega < 1 \): Purifying (negative) selection
- \( \omega > 1 \): Positive (diversifying) selection

---

The selection ratio test:

Comparison of the rate of Nonsynonymous (replacement) substitutions (dN, \( K_A \)) to Synonymous (silent) substitutions (dS, \( K_S \))

Higher dN/dS ratios can be produced by
- positive selection favoring a change in function
- relaxation of purifying selection (rapid increase in population size)

Rules of thumb:

<table>
<thead>
<tr>
<th>dN/dS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1 - 0.2</td>
<td>no evidence of selection</td>
</tr>
<tr>
<td>0.2 - 1</td>
<td>can’t distinguish between positive election and relaxed selection</td>
</tr>
<tr>
<td>&gt; 1</td>
<td>positive selection</td>
</tr>
</tbody>
</table>

ARE WE DONE????
The dN/dS ratio test: applications

\[ \text{Test} = \frac{dN}{dS} \]

where:

- \( dS \): synonymous/silent rate of change
- \( dN \): nonsynonymous rate of change

Substitutions / Silent Site (KS)
Substitutions / Replacement Site (KA)

Open points denote genes for which the ratio dN/dS is not significantly different from 1.

Progressive decline of $dN/dS$ indicates gradual strengthening of ‘purifying’ selection.

$dN$ (replacement)

$dS$ (synonymous, silent)


Great apes
Old World monkey
Rodents
Artiodactyls

Evolution of leptin

Benner et al. 2002 Science 296:864
Alas, raw counting of dN, dS, doesn’t work: all sites are not created equal...

Example:

<table>
<thead>
<tr>
<th>Sequence 1</th>
<th>Asp</th>
<th>Thr</th>
<th>Ala</th>
<th>Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>AC</td>
<td>CG</td>
<td>GTT</td>
<td></td>
</tr>
<tr>
<td>Sequence 2</td>
<td>CG</td>
<td>AC</td>
<td>CG</td>
<td>GTT</td>
</tr>
<tr>
<td>Ala</td>
<td>Thr</td>
<td>Ser</td>
<td>Val</td>
<td></td>
</tr>
</tbody>
</table>

\[ \bar{dN} = 2 \]
\[ dS = 1 \]
\[ dN/dS = 2 = \text{diversifying selection? No!} \]

We must calculate dN and dS wrt the possible number of nonsynonymous and synonymous sites, respectively.

Complications with simple counting: redundancy in the codon table means we have to weight each site.

<table>
<thead>
<tr>
<th>First base (5′ end)</th>
<th>Second base</th>
<th>Third base (3′ end)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UU</td>
<td>Phe</td>
<td>Leu</td>
</tr>
<tr>
<td>UUC</td>
<td>UCU</td>
<td>UCG</td>
</tr>
<tr>
<td>UUA</td>
<td>UAU</td>
<td>UAC</td>
</tr>
<tr>
<td>UUG</td>
<td>UUA</td>
<td>UAG</td>
</tr>
<tr>
<td>CUU</td>
<td>Leu</td>
<td>CCA</td>
</tr>
<tr>
<td>CUC</td>
<td>GAG</td>
<td>GGG</td>
</tr>
<tr>
<td>CUA</td>
<td>AAA</td>
<td>Arg</td>
</tr>
<tr>
<td>AUG</td>
<td>Met</td>
<td>UAA</td>
</tr>
<tr>
<td>AUU</td>
<td>ACC</td>
<td>Thr</td>
</tr>
<tr>
<td>AUC</td>
<td>ACA</td>
<td>Arg</td>
</tr>
<tr>
<td>AUA</td>
<td>AUC</td>
<td>Thr</td>
</tr>
<tr>
<td>AUG</td>
<td>Met</td>
<td>UAA</td>
</tr>
<tr>
<td>GIU</td>
<td>GAG</td>
<td>Arg</td>
</tr>
<tr>
<td>GIU</td>
<td>GAG</td>
<td>Arg</td>
</tr>
<tr>
<td>GIU</td>
<td>GAG</td>
<td>Arg</td>
</tr>
<tr>
<td>GIU</td>
<td>GAG</td>
<td>Arg</td>
</tr>
<tr>
<td>GIU</td>
<td>GAG</td>
<td>Arg</td>
</tr>
<tr>
<td>GIU</td>
<td>GAG</td>
<td>Arg</td>
</tr>
<tr>
<td>GIU</td>
<td>GAG</td>
<td>Arg</td>
</tr>
<tr>
<td>GIU</td>
<td>GAG</td>
<td>Arg</td>
</tr>
</tbody>
</table>

THR is 4-fold degenerate
ASP is 2-fold degenerate
### Synonymous positions calculation

Example:

<table>
<thead>
<tr>
<th>Degeneracy</th>
<th>Asp</th>
<th>Thr</th>
<th>Ala</th>
<th>Val</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>112</td>
<td>114</td>
<td>114</td>
<td>114</td>
</tr>
</tbody>
</table>

**Sequence 1**

<table>
<thead>
<tr>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU</td>
<td>UCU</td>
<td>UCA</td>
<td>UCG</td>
</tr>
<tr>
<td>UUC</td>
<td>UCC</td>
<td>UCA</td>
<td>UCG</td>
</tr>
<tr>
<td>UUA</td>
<td>UCU</td>
<td>UCA</td>
<td>UCG</td>
</tr>
<tr>
<td>UUG</td>
<td>UCG</td>
<td>UCA</td>
<td>UCG</td>
</tr>
</tbody>
</table>

Q: How many possible *synonymous nt positions* are there in sequence 1?

A: The 0-fold (nondegenerate) positions can never be synonymous - they count 0, because any nt change results in a different amino acid. There is 1 two-fold degenerate position (the last position for Asp), and there are 3 four-fold degenerate positions (for Thr and Ala and Val).

Q: How much does a two-fold site count?

A: 1/3 (one out of 3 changes are synonymous)

---

### Synonymous positions calculation

Example:

<table>
<thead>
<tr>
<th>Degeneracy</th>
<th>Asp</th>
<th>Thr</th>
<th>Ala</th>
<th>Val</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>112</td>
<td>114</td>
<td>114</td>
<td>114</td>
</tr>
</tbody>
</table>

**Sequence 1**

<table>
<thead>
<tr>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU</td>
<td>UCU</td>
<td>UCA</td>
<td>UCG</td>
</tr>
<tr>
<td>UUC</td>
<td>UCC</td>
<td>UCA</td>
<td>UCG</td>
</tr>
<tr>
<td>UUA</td>
<td>UCU</td>
<td>UCA</td>
<td>UCG</td>
</tr>
<tr>
<td>UUG</td>
<td>UCG</td>
<td>UCA</td>
<td>UCG</td>
</tr>
</tbody>
</table>

- # total of *synonymous* positions = 3 (for the three 4-fold sites) + 0.33 (for the single 2-fold position) + 0 (for the 8 nondegenerate positions) = **3.33**
# Nonsynonymous positions calculation

Example:

| Degeneracy | 112 114 114 114 |
| Asp          | Thr  Ala  Val |
| Sequence 1   | GAC  ACA  GCG  GTT |

Q: How many nonsynonymous positions are there in sequence 1?

A: 8 nondegenerate positions + 2/3 for the 2-fold position = 8.67

Example:

| Degeneracy 1 | 112 114 114 114 114 |
| Asp          | Thr  Ala  Val |
| Sequence 1   | GAC  ACA  GCG  GTT |
| Sequence 2   | GCC  ACT  TCG  GTT |
| Ala          | Thr  Ser  Val |

Sequence 2 has 8 nonsynonymous sites and 4 synonymous sites
(check this)

For the dN/dS comparison, we average the numbers from both sequences
(Why?)

# Nonsynonymous positions = (8.66 + 8)/2 = 8.33
# Synonymous positions = (3.33 + 4)/2 = 3.67

Finally, back to the sequence comparison...

Example:

| Degeneracy 1 | 112 114 114 114 114 |
| Asp          | Thr  Ala  Val |
| Sequence 1   | GAC  ACA  GCG  GTT |
| Sequence 2   | GCC  ACT  TCG  GTT |
| Ala          | Thr  Ser  Val |

| Degeneracy 2 | 114 114 114 114 |

Sequence 2 has 8 nonsynonymous sites and 4 synonymous sites
(check this)
Example:

<table>
<thead>
<tr>
<th>Degeneracy 1</th>
<th>112</th>
<th>114</th>
<th>114</th>
<th>114</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asp</td>
<td>Thr</td>
<td>Ala</td>
<td>Val</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence 1</th>
<th>GAC</th>
<th>ACA</th>
<th>GCG</th>
<th>GTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence 2</td>
<td>GCC</td>
<td>ACT</td>
<td>TCG</td>
<td>GTT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Degeneracy 2</th>
<th>114</th>
<th>114</th>
<th>114</th>
<th>114</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>Thr</td>
<td>Ser</td>
<td>Val</td>
<td></td>
</tr>
</tbody>
</table>

There are 2 nonsynonymous changes,
So $dN = 2/8.33 = 0.24$

There is 1 silent change,
So $dS = 1/3.67 = 0.27$

$dN/dS = 0.23/0.27$, or $\omega = 0.88 < 1$ despite having more nonsynonymous changes than synonymous change no evidence of positive selection (compare to our earlier calculation of 2.0 as the ratio earlier)

Example application: HIV analysis

Note: the Y-axis is labeled with the inverse ratio $dS/dN$ rather than $dN/dS$ - this is typical!

- HIV-1 *env* sequences sampled from a seroconverted hemophiliac patient 3, 4, 5, 6 & 7 years after infection.
- For each year they took the mean $dS$ and mean $dN$ derived from all possible pairwise combinations.
- At year 3 the $dN/dS$ ratio is $10^{-1}$, indicating strong positive selection but as infection progresses the selection pressure declines noticeably.
- Early in infection the immune system responds against common virus variants providing a strong positive selection pressure for diversification.
- As the immune system declines the selection pressure becomes weaker.

**Figure:**

Many studies have used dN/dS to detect selection

<table>
<thead>
<tr>
<th>Table 1. Selected examples of protein-coding genes in which positive selection was detected by using the dN/dS ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Class I chemokine gene</td>
</tr>
<tr>
<td>CT10 genes</td>
</tr>
<tr>
<td>Defensin genes</td>
</tr>
<tr>
<td>IFN-γ genes</td>
</tr>
<tr>
<td>Immunoglobulin V, J, λ genes</td>
</tr>
<tr>
<td>MHC class I genes</td>
</tr>
<tr>
<td>Polygalacturonase inhibitor 1 gene</td>
</tr>
<tr>
<td>Toll-like receptor genes</td>
</tr>
<tr>
<td>Abelson myeloid leukemia virus genes</td>
</tr>
<tr>
<td>Transformin genes</td>
</tr>
<tr>
<td>Urokinase-type plasminogen activator genes</td>
</tr>
<tr>
<td>α1-Proteinase inhibitor genes</td>
</tr>
<tr>
<td>Genes involved in neutralizing systems or immunity</td>
</tr>
<tr>
<td>Rapid gene</td>
</tr>
<tr>
<td>C2, TRAP, USA-1 and RPSE</td>
</tr>
<tr>
<td>Delta-1 antigen gene</td>
</tr>
<tr>
<td>E. coli</td>
</tr>
<tr>
<td>Envelope gene</td>
</tr>
<tr>
<td>G protein-coupled receptor genes</td>
</tr>
<tr>
<td>Genes involved in innate immunity</td>
</tr>
<tr>
<td>Cathepsin D gene</td>
</tr>
<tr>
<td>Endothelial progenitor cell genes</td>
</tr>
<tr>
<td>Polygalacturonase genes</td>
</tr>
<tr>
<td>Nicotinamide adenine dinucleotide</td>
</tr>
<tr>
<td>Polycomb group genes</td>
</tr>
<tr>
<td>Silencer factor 1 gene</td>
</tr>
<tr>
<td>Signaling pathway genes</td>
</tr>
<tr>
<td>Survival determinant gene</td>
</tr>
</tbody>
</table>

So, are we done now?

I wish! There are at least two more sets of corrections to consider… to accurately measure the # of evolutionary events between sequence 1 and sequence 2…

What are the assumptions we have tacitly assumed?

Assumption #1 (uniformity): the pr of one nucleotide changing into another is the same for all nucleotides.

e.g., pr A→C is the same as pr A→G, T→G, ...

But this is not true!

Assumption #2 (parsimony): The observed sequence differences are due to the fewest possible # of changes.

e.g., if we observe A→C, then this was not the result of, say, A→T→C, or A→T→C→A→T→C, etc.
In short...2 basic problem areas

- The codon change (evolution) problem
- The multiple path/parsimony problem

We can pose these ‘issues’ together in this way:

ATG CGC GTG TTA GAA
. . . . . G . C . . . . .
ACG ATT GTT GCT TTA . .
. . . . . . C . A . . . . . .

What is $K$, the actual number of substitutions, given the sequence length $N$ and the # of observed differences, $D$?
So how do you estimate $dN$ or $dS$?

Just as complicated as estimating nucleotide substitutions!
When species are close, you can just use mutations
But, when they are farther apart, multiple substitutions can occur in the same position and the number of changes is *underestimated* by simple counting (what would plot of *observed* vs. *estimated substitutions* look like?)

Calculation of $dN$ and $dS$ requires quite complex models of nucleotide substitution
The Parsimony problem (see Lewontin, 1989)

There were no guarantees that a particular site had not undergone multiple changes. Two possible scenarios where multiple substitutions at a single site would lead to underestimation of the number of substitutions that had occurred if a simple count were performed.

Solution to the nt substitution problem, and, partly, to the multiple substitution problem

Idea: given some observed pattern of nt substitutions, how do we match it with an evolutionary model?

This leads to progressively more sophisticated (parameterized) models for nt substitution...
Codon models through the ages: These are all variants of the following general Markov model.

Modeling nucleotide changes

The simplest correction: Jukes-Cantor (1969)

\[
\begin{align*}
\frac{dD}{dK} &= \frac{N-D}{N} - \frac{1}{3N} \\
\frac{1}{1-\frac{4D}{3N}} &= dK \\
\frac{K}{N} &= -\frac{3}{4} \ln \left(1 - \frac{4D}{3N}\right)
\end{align*}
\]
J-K model

- From this model, the overall rate of substitution for any given nucleotide is $3\alpha$.
- If a site within a gene was occupied by a C at time 0, then the probability ($p$) that this site would still be the same nucleotide at time 1 would be
  $$p_{C(1)} = 1 - 3\alpha$$
- Because a reversion to C could occur if the original C changed to another nucleotide in that first time span, at time 2 the probability would be:
  $$p_{C(2)} = (1 - 3\alpha)p_{C(1)} + \alpha[1 - p_{A(1)}]$$

Temporal change in the probability of having a certain nucleotide, say A, at a given nucleotide site

\[
\begin{align*}
p_{A(0)} &= \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \\
p_{A(t)} &= \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}
\end{align*}
\]
Jukes-Cantor Model: correcting counted nt differences for multiple hits

• If we invert the J-K formula, we can derive the equation that yields an estimate of the true number of substitutions that have occurred between two sequences when only a pairwise counting of differences is observable:

\[ K = -\frac{3}{4}ln[1 - \left(\frac{4}{3}\right)p] \]

• Where K is the actual number of substitutions per site, and p is the fraction of nucleotides that a simple count reveals to be different between the two sequences.

Transitions vs. transversions

\[ K = \text{transition/transversion ratio} \]
The need for more complex models

More transitions than transversions are synonymous.

Nearly 50% of synonymous changes are transitions while less than 30% of nonsynonymous changes are transitions.

(prove this!)
Markov chain model: derivation of Jukes-Cantor
(next 4 slides - optional)

State space = \{A,C,G,T\}.

\[ p(i,j) = \text{pr}(\text{next state } S_j \mid \text{current state } S_i) \]

Markov assumption:
\[ p(i,j) = \text{pr}(\text{next state } S_j \mid \text{current state } S_i \& \text{any configuration of states before this}) \]

Only the present state, not previous states, affects the probs of moving to next states.
\( r_{ij} = \text{rate of substitution from state } i \text{ to state } j \)
\( p(i,j) = \text{pr of being in state } j \text{ at time } t \)
\( \text{given that ancestor was in state } i \text{ at time 0} \)

The multiplication rule

\[
pr(\text{state after next is } S_k \mid \text{current state is } S_i) = \sum_j pr(\text{state after next is } S_k, \text{next state is } S_j \mid \text{current state is } S_i) \tag{addition rule}
\]

\[
= \sum_j pr(\text{next state is } S_j \mid \text{current state is } S_i) \times pr(\text{state after next is } S_k \mid \text{current state is } S_i, \text{next state is } S_j) \tag{multiplication rule}
\]

\[
= \sum_j p(i,j) \times p(j,k) \tag{Markov assumption}
\]

\[
= (i,k)\text{-element of } P^2, \text{ where } P=(p(i,j)).
\]

More generally,

\[
pr(\text{state } t \text{ steps from now is } S_k \mid \text{current state is } S_i) = i,k\text{ element of } P^t
\]
First cut: Jukes-Cantor model (1969)

All base frequencies = 1/4; all substitution rates = \( \alpha \)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>( \alpha )</td>
<td>( \alpha )</td>
<td>( \alpha )</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>( \alpha )</td>
<td>( \alpha )</td>
<td>( \alpha )</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>( \alpha )</td>
<td>( \alpha )</td>
<td>( \alpha )</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>( \alpha )</td>
<td>( \alpha )</td>
<td>( \alpha )</td>
<td></td>
</tr>
</tbody>
</table>

Diagram of the Jukes-Cantor model of nucleotide substitution. For their model, Jukes and Cantor assumed that all nucleotides changed to each of the three alternative nucleotides at the same rate, \( \alpha \).

Let \( p(t) \) be the probability of a particular base at time \( t \) being a particular base.

After elapsed time \( dt \), mutation to other three bases reduces that by \( -3 \alpha dt \).

While the gain from other bases is \( \alpha dt(1-p(t)) \)

So: \( p(t+dt) = p(t) - 3 \alpha dt + \alpha dt(1-p(t)) \) or \( dp(t)/dt = -4 \alpha p(t) \)

Write \( p(t) = a e^{(-b t)} + c \), solution is \( b = 4 \alpha \), \( c = 1/4 \) or \( p(t) = a e^{(-4 \alpha t)} + 1/4 \)

If \( p(0) = 1 \), then \( \alpha = 3/4 \)

If \( p(0) = 0 \), then \( \alpha = -1/4 \) so we have:

\[
\begin{align*}
Psame &= 1/4 + 3/4 e^{(-4 \alpha t)} \\
Pchange &= 1/4 - 1/4 e^{(-4 \alpha t)}
\end{align*}
\]
Jukes-Cantor model results

\[ p_{ii}(t) = \frac{1}{4} + \frac{3}{4} \exp(-4\alpha t) \]
\[ p_{ij}(t) = \frac{1}{4} - \frac{1}{4} \exp(-4\alpha t) \]
\[ \text{pr(different)} = 1 - p_{ii}(2t) = \frac{3}{4} (1 - \exp(-8\alpha t)) \]
\[ = 3p_{ij}(2t) \]

Temporal change in the probability of having a certain nucleotide, say A, at a given nucleotide site.

\[ \alpha = 5 \times 10^{-9} \]

Jukes Cantor
Illustration of J-K correction

1. Human and bovine beta-globins are aligned with no deletions at 145 out of 147 sites. They differ at 23 of these sites. Thus $n_o/n = 23/145$, and the corrected distance using the Jukes-Cantor formula is (natural logs)

   $$- \frac{19}{20} \times \log \left(1 - \frac{20}{19} \times \frac{23}{145}\right) = 17.3 \times 10^{-2}.$$ 

2. The human and gorilla sequences are aligned without gaps across all 300 bp, and differ at 14 sites. Thus $n_o/n = 14/300$, and the corrected distance using the Jukes-Cantor formula is

   $$- \frac{3}{4} \times \log \left(1 - \frac{4}{3} \times \frac{14}{300}\right) = 4.8 \times 10^{-2}.$$ 

Kimura’s Two-Parameter Model

Diagram of Kimura’s two-parameter model of nucleotide substitution. Kimura assumed that nucleotide substitutions occurred at essentially two different rates: $\alpha$ for transitions, and $\beta$ for transversions.
J-K: one parameter, alpha (= equal rate of change to some other nt)
Kimura: two parameters, alpha & beta (1 rate for transversions, one rate for transitions)

From Jukes-Cantor to Kimura 2 parameter model
Kimura 2 parameter model: assumes rate of change different for transitions vs. transversions

\[ P_{JC} = \begin{bmatrix} \alpha & \alpha & \alpha \\ \alpha & \alpha & \alpha \\ \alpha & \alpha & \alpha \end{bmatrix} \]
\[ f = \left[\frac{1}{4}\right] \]
\[ \kappa = \left(\frac{1}{4}\right) \left(1 - \frac{4p}{3}\right) \]

\[ P_{K2} = \begin{bmatrix} \beta & \alpha & \beta \\ \beta & \beta & \beta \\ \alpha & \beta & \alpha \\ \alpha & \beta & \alpha \end{bmatrix} \]
\[ f = \left[\frac{1}{4} \frac{1}{4} \frac{1}{4} \frac{1}{4}\right] \]

\[ K = \frac{1}{2} \ln \left(1/(1 - 2p - 0)\right) + \frac{1}{2} \ln \left(1/(1 - 0)\right) \]
Kimura’s Two-Parameter Model (K2P)

- The true number, $K$, of substitutions that have occurred between two sequences when only a pairwise counting of differences is available:

$$K = -3 \ln[1 - (4/3)(p)]$$

Where $p$ is the fraction of nucleotides that a simple count reveals to be transitions, and $q$ is the fraction that are transversions.

Compare J-K: $K = \frac{1}{2} \ln\left[\frac{1}{1 - 2p - q}\right] + \frac{1}{4} \ln\left[\frac{1}{1/2q}\right]$

So much for the unequal nt change issue...what about the multiple paths problem?
Solution to the parsimony problem

To estimate the actual number of synonymous and non-synonymous differences between a pair of sequences we must take account of every possible path between the pairs of codons being compared

1 difference → 1 path
2 differences → 2 paths
3 differences → 6 paths

(the real Solution: maximum likelihood - test out all the paths)

Nei-Gojoburi method: take into account all possible such paths, just as you have just seen for this one example

- Consider all possible pathways between each pair of homologous codons
- dS, dN are the synonymous, nonsynonymous substitutions as usual
- For GTT (Val) and GTA(Val), there is 1 synon and 0 nonsynon differences: dS and dN are 1 and 0
Detecting natural selection by measuring substitution rates is one of the most fruitful methods of detecting the historical activity of natural selection. To compare the relative rates of synonymous (dS) and nonsynonymous (dN) nucleotide substitutions:

- **Negative or purifying selection** has occurred if \( ds > dn \).
- **Positive or Darwinian selection** has occurred if \( ds < dn \).
- Neutral conditions have prevailed when \( ds = dn \).

But how do you estimate dS and dN? It's just as complicated as estimating nucleotide substitutions - many approaches and few clear ideas about which is most efficient and accurate (and when):

- **Evolutionary methods**: examines all possible evolutionary pathways between two sequences and averages rates.
- **K2P-basted methods**: takes advantage of K2P's model and classifies sites as 4-old, 2-old, and 0-old degenerate; more accurate than evolutionary methods.
- **MLE methods**: based on HKY model, very robust statistically, most commonly used.

**Revised protocol (Nei and Gojoburi, NG86):**

1. Count the number of synonymous (\( S_i \)) and non-synonymous (\( N_i \)) sites (using fractions as appropriate) in each sequence.
2. \( N = (N_1 + N_2) / 2 \); \( S = (S_1 + S_2) / 2 \)
3. Count the number of synonymous and non-synonymous differences by summing over all possible paths between each pair of codons.
4. Correct for multiple substitutions (using Jukes and Cantor; or Kimura; or your favorite correction method).
5. Test significance using a Z-test (note: this is a large-sample statistical test – c.f. central limit theorem) or Fisher's Exact Test for comparison of within species vs. outside species dN/dS values (this is the McDonald-Kreitman test).

**Consider:**

There are 6 pathways between TTG and AGA (each with 3 changes):

\[ \text{frog: TTG} \rightarrow \text{dog: AGA} \]

1. TTG(Leu) → AGG(Arg) → AGA(Arg); \( 1,2 \)
2. TTG(Leu) → AGT(Met) → AGA(Arg); \( 1,3 \)
3. TTG(Leu) → TGG(Trp) → AGG(Arg) → AGA(Arg); \( 1,2 \)
4. TTG(Leu) → TGG(Trp) → TGA(Ter) → AGA(Arg); *
5. TTG(Leu) → TTA(Leu) → ATA(Ile) → AGA(Arg); \( 1,2 \)
6. TTG(Leu) → TTA(Leu) → TGA(Ter) → AGA(Arg); *

We can ignore 4 & 6; \( S = 3/4 \) & \( N = 9/4 \) (\( 3/4 + 9/4 = 3 \))
Statistical tests

- For large samples, use $Z = \frac{D}{\sigma(D)}$
- where $D = dN - dS$
- $Z$ is the z-test score
- The test is usually one-tailed, since we only care if $dN > dS$

Taking stock:
From pairwise counting to likelihood models

Pairwise methods: simple, intuitive, popular
But have complications in their estimates...
Step 1. Count $N$ and $S$ sites
Step 2. Count $N$ and $S$ differences
Step 3. Apply some correction (for multiple hits e.g.)
Counting methods are SAFE to use if
codon usage, esp. at 3rd position, is \textbf{uniform},
the sequences are not very divergent, and
transition/transversion rates, $K$, are similar

But what if they are not???

![Selection varies in space (where)](image)

Fig. 1. The identification of sites under positive selection from the sperm lysozyme genes of 25
salmonid species. (a) Posterior probabilities for site classes with different $w$ ratios along the
sequence. The lysozyme sequence of the redsalmonid \textit{Oncorhynchus} is shown before the area.
 Millions of years (My) estimates under Model 2 (blockP) suggest three site classes with the $w$ ratios at
$w_9 = 0.285$ (grey), $w_9 = 0.911$ (green) and $w_9 = 3.095$ (blue), and with proportions $p_1 = 0.329$,
$p_2 = 0.420$ and $p_3 = 0.259$. 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 after
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, 0.095 and
0.065, and thus the site is likely to be under strong purifying selection. The posterior probabilities
for Site 4 are 0.003, 0.001 and 1.000, and thus this site is almost certainly under diversifying
selection. 3D Lysozyme structure from the redsalmonid (Protein Data Bank file 1OZQ), with sites
under positive selection marked with red and blue. The N terminus is the amino acid sequence
for olive flounder, because the first three amino acids are unisolated. The five $w$
ratios are indicated: $w_1$ from amino acids 73 to 83, $w_2$ from 84 to 94, $w_3$ from 95 to 105, $w_4$
from 96 to 107 and $w_5$ 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 crys-
tal structure. Reproduced with permission, from Ref. 20.
Selection pressure varies in time

40 – 80 mya

150 – 200 mya

100 – 140 mya

35 mya

Chrom. 11

β globin gene cluster

Effect of kappa on $S$

$S$ when kappa is estimated

$S$ when codon frequencies are empirically estimated

(Data from: Bielawski, Dunn, and Yang (2000) Genetics: 156: 1299-1308)
Simulation study of bias in dS

(Data from: Dunn, Bielawski, and Yang (2001) Genetics, 157: 295-305)

Why does this matter?

Estimation of $d_S$ and $d_N$ between Drosophila melanogater and D. simulans GstD1 genes

<table>
<thead>
<tr>
<th>Method</th>
<th>ts/tv bias</th>
<th>Codon bias</th>
<th>$K$</th>
<th>$S$</th>
<th>$N$</th>
<th>$d_S$</th>
<th>$d_N$</th>
<th>$\omega$</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>no</td>
<td>1.0</td>
<td>162.9</td>
<td>447.1</td>
<td>0.0778</td>
<td>0.0213</td>
<td>0.274</td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>1.88</td>
<td>165.8</td>
<td>434.2</td>
<td>0.0221</td>
<td>0.0691</td>
<td>0.320</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>3 × 4</td>
<td>1.0</td>
<td>70.6</td>
<td>529.4</td>
<td>0.1605</td>
<td>0.0189</td>
<td>0.118</td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>3 × 4</td>
<td>2.71</td>
<td>73.4</td>
<td>526.6</td>
<td>0.1526</td>
<td>0.0193</td>
<td>0.127</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>empirical</td>
<td>1.0</td>
<td>40.5</td>
<td>559.5</td>
<td>0.3198</td>
<td>0.0201</td>
<td>0.063</td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>empirical</td>
<td>2.53</td>
<td>45.2</td>
<td>554.8</td>
<td>0.3041</td>
<td>0.0204</td>
<td>0.067</td>
<td></td>
</tr>
</tbody>
</table>

Why this matters

human-chimp mt Nδ5 gene
(known Tv/Ts bias)

<table>
<thead>
<tr>
<th>method</th>
<th>d_S</th>
<th>d_N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nei-Gojobori (K=0.5/no bias)</td>
<td>41.94 ± 3.80</td>
<td>3.79 ± 0.54</td>
</tr>
<tr>
<td>Li-Wu-Luo (K=0.5/no bias)</td>
<td>42.77 ± 4.14</td>
<td>3.78 ± 0.54</td>
</tr>
<tr>
<td>modified NK (K=0.20)</td>
<td>27.90 ± 2.40</td>
<td>4.38 ± 0.62</td>
</tr>
<tr>
<td>Goldman-Yang (ω computed)</td>
<td>28.72</td>
<td>4.42</td>
</tr>
</tbody>
</table>

Values multiplied by 100

Key points

1. Assumptions matter more than methods (usually).
2. Ignoring the transition/transversion rate bias leads to underestimation of S, overestimation of d_S and underestimation of the d_S/d_S (ω) ratio.
3. Codon-usage bias often has the opposite effect to the transition/transversion bias and can be more important.
4. Different assumptions can produce different estimates even when the sequences are highly similar.
5. Different assumptions can lead to different biological conclusions.
General solution: maximum likelihood methods

How to fix: Maximum likelihood method

The codon is considered the unit of evolution. The substitution rate from codons $i$ to $j$ ($i \neq j$) is given as:

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

Parameter $\kappa$ is the transition/transversion rate ratio, $\pi_i$ is the equilibrium frequency of codon $j$ and $\omega = \frac{\pi_{j,1}}{\pi_{j,0}}$ measures the selective pressure on the protein. The $q_i$ are relative rates because time and rate are confounded in such an analysis. Given the rate matrix $Q = \{q_i\}$, the transition probability matrix over time $t$ is calculated as:

$$P(t) = (p_i(t)) = e^{tQ}$$

where $p_i(t)$ is the probability that codon $i$ becomes codon $j$ after time $t$. Likelihood calculation on a phylogeny involves summing over all possible codons in extinct ancestors (internal nodes of the tree).
Advantages of Maximum likelihood method

1. We can take advantage of the phylogeny
2. Computation of transition probabilities accomplishes the following in 1 step:
   i. estimation of parameters (t, κ, ω)
   ii. correction for multiple hits
   iii. weight evolutionary pathways between codons

Important parameters:

• Transition/transversion rate ratio: κ
• Biased codon usage: π for codon j
• Nonsynonymous/synonymous rate ratio: ω = dN/dS
Rate matrix $Q$: basic Markov transition matrix for all possible changes

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

$$P(t) = (p_i(t)) = \Phi^Q$$

(Goldman & Yang 1994 MBE 11:725-736)

Maximum likelihood method

- Uses general rate matrix $Q$
- Adds to this, computation over all possible paths from ancestral state to current observed nucleotides
How to get to CTG

Synonymous
\[ \text{CTC (Leu)} \rightarrow \text{CTG (Leu)}: \pi_{\text{CTG}} \]
\[ \text{TTC (Leu)} \rightarrow \text{CTG (Leu)}: \kappa \pi_{\text{CTG}} \]

Nonsynonymous
\[ \text{GTG (Val)} \rightarrow \text{CTG (Leu)}: \omega \pi_{\text{CTG}} \]
\[ \text{CCG (Pro)} \rightarrow \text{CTG (Leu)}: \kappa \omega \pi_{\text{CTG}} \]

Simplest case: two codons

\[ L_x(CCC, CCT) = \sum_i \pi_x p_{CCC}(t_i) p_{CCT}(t_i) \]

Note: analysis is typically done by using an unrooted tree

\( pr \) observing a site with codons \( i \) and \( j \) in the two sequences is just \( \pi_{p_{ij}}(t) \)
Sum over all possible codons for *each* ancestral node

The likelihood of observing the entire sequence alignment is the product of the probabilities at each site.

$$L = L_1 \times L_2 \times \cdots \times L_N = \prod_{h=1}^{N} L_h$$

The log likelihood is a sum over all sites.

$$\ell(t, \kappa, \omega) = \ln(L) = \ln(L_1) + \ln(L_2) + \ln(L_3) + \cdots + \ln(L_N) = \sum_{h=1}^{N} \ln(L_h)$$

Remember: we are interested in adaptive evolution

$$\omega = 1: \text{neutral evolution}$$
$$\omega < 1: \text{purifying (negative) selection}$$
$$\omega > 1: \text{diversifying (positive) selection}$$
Example comparison of methods

Table 2. Estimation of $d_s$ and $d_N$ between the human and orangutan $\beta$-globin genes (142 codons)$^a$

<table>
<thead>
<tr>
<th>Method and/or model</th>
<th>$x$</th>
<th>$S$</th>
<th>$N$</th>
<th>$d_s$</th>
<th>$d_N$</th>
<th>$d_s/d_N$</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approximate methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li &amp; Li</td>
<td>3.9</td>
<td>137.1</td>
<td>389.9</td>
<td>0.0030</td>
<td>0.0024</td>
<td>0.109</td>
<td>16</td>
</tr>
<tr>
<td>(F) $F_{1/4}$, $w$ = 1 fixed</td>
<td>1.0</td>
<td>129.1</td>
<td>389.9</td>
<td>0.0032</td>
<td>0.0029</td>
<td>0.137</td>
<td>610.40</td>
</tr>
<tr>
<td>(F) $F_{1/4}$, $w$ estimated</td>
<td>1.0</td>
<td>137.1</td>
<td>389.9</td>
<td>0.0030</td>
<td>0.0024</td>
<td>0.109</td>
<td>610.40</td>
</tr>
<tr>
<td>(F) $F_{1/4}$, $w$ = 1 fixed</td>
<td>1.0</td>
<td>129.1</td>
<td>389.9</td>
<td>0.0032</td>
<td>0.0029</td>
<td>0.137</td>
<td>610.40</td>
</tr>
<tr>
<td>(F) $F_{1/4}$, $w$ estimated</td>
<td>1.0</td>
<td>137.1</td>
<td>389.9</td>
<td>0.0030</td>
<td>0.0024</td>
<td>0.109</td>
<td>610.40</td>
</tr>
<tr>
<td>(F) $F_{1/4}$, $w$ = 1 fixed</td>
<td>1.0</td>
<td>129.1</td>
<td>389.9</td>
<td>0.0032</td>
<td>0.0029</td>
<td>0.137</td>
<td>610.40</td>
</tr>
<tr>
<td>(F) $F_{1/4}$, $w$ estimated</td>
<td>1.0</td>
<td>137.1</td>
<td>389.9</td>
<td>0.0030</td>
<td>0.0024</td>
<td>0.109</td>
<td>610.40</td>
</tr>
<tr>
<td>(F) $F_{1/4}$, $w$ = 1 fixed</td>
<td>1.0</td>
<td>129.1</td>
<td>389.9</td>
<td>0.0032</td>
<td>0.0029</td>
<td>0.137</td>
<td>610.40</td>
</tr>
<tr>
<td>(F) $F_{1/4}$, $w$ estimated</td>
<td>1.0</td>
<td>137.1</td>
<td>389.9</td>
<td>0.0030</td>
<td>0.0024</td>
<td>0.109</td>
<td>610.40</td>
</tr>
<tr>
<td>(F) $F_{1/4}$, $w$ = 1 fixed</td>
<td>1.0</td>
<td>129.1</td>
<td>389.9</td>
<td>0.0032</td>
<td>0.0029</td>
<td>0.137</td>
<td>610.40</td>
</tr>
<tr>
<td>(F) $F_{1/4}$, $w$ estimated</td>
<td>1.0</td>
<td>137.1</td>
<td>389.9</td>
<td>0.0030</td>
<td>0.0024</td>
<td>0.109</td>
<td>610.40</td>
</tr>
<tr>
<td>(F) $F_{1/4}$, $w$ = 1 fixed</td>
<td>1.0</td>
<td>129.1</td>
<td>389.9</td>
<td>0.0032</td>
<td>0.0029</td>
<td>0.137</td>
<td>610.40</td>
</tr>
<tr>
<td>(F) $F_{1/4}$, $w$ estimated</td>
<td>1.0</td>
<td>137.1</td>
<td>389.9</td>
<td>0.0030</td>
<td>0.0024</td>
<td>0.109</td>
<td>610.40</td>
</tr>
</tbody>
</table>

Matlab MBE package includes NG86 (Nei-Gojobari method, a counting method), and GY94 (a maximum likelihood method)

Task 1: software to estimate $d_s$ and $d_N$

<table>
<thead>
<tr>
<th>Methods</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting methods</td>
<td>MEGA, codeml &amp; ym00 in PAML</td>
</tr>
<tr>
<td>NG86</td>
<td>DAMBE</td>
</tr>
<tr>
<td>L93</td>
<td>DIVERGE by Comeron, ym00 in PAML</td>
</tr>
<tr>
<td>Comeron 95</td>
<td></td>
</tr>
<tr>
<td>YN00</td>
<td></td>
</tr>
<tr>
<td>ML methods</td>
<td>codeml in PAML</td>
</tr>
<tr>
<td>GY94</td>
<td>HPHY</td>
</tr>
<tr>
<td>MG94</td>
<td></td>
</tr>
</tbody>
</table>

---

$^a$GenBank accession numbers are V00rAQ (human) and MAPAr– (orangutan).
$^b$Frequencies equal codon frequencies at three codon positions are used to calculate codon frequencies (3 free parameters). $^c$Codon frequencies at three codon positions are used to calculate codon frequencies (3 free parameters). $^d$Codon frequencies are used as free parameters (30 free parameters).
$^e$The log likelihood value.
ML estimate of parameter values

\[ t, \kappa, \omega = \text{unknown values} \]

\[ \pi's = \text{empirical [F3×4 or F61]} \]

Use numerical hill-climbing algorithm to maximize the likelihood function

MLE of \( t \) and \( \omega \), with \( \kappa \) fixed

- codeml package in PAML; also in MBE package in MATLAB
ML estimation of dN and dS

Number of substitutions are calculated from $q$ and $t$
Number of sites ($S$ and $N$) are calculated from $q$ by fixing $\omega = 1$

Remaining problem...

In a pairwise analysis we must average the $\omega$ ratio over:
1. all sites
2. the entire evolutionary history

In a large-scale pairwise database search, only 17 out of 3,595 genes were found to be under positive selection, at <0.5% (Endo et al. 1996 MBE 13: 685-690)
Solution: use the likelihood ratio test (LRT) to compare hypotheses

\[ \ell_0 \text{ is the maximum log likelihood under } H_0 \text{ with parameters } \theta_0 \]
and
\[ \ell_1 \text{ is the maximum log likelihood under } H_1 \text{ with parameters } \theta_1 \]

Test statistic \[ 2\Delta \ell = 2(\ell_0(\theta_0) - \ell_1(\theta_1)) \]

This test statistic will (usually) be distributed as \( \chi^2 \)

Degrees of freedom = difference in the number of parameters between The two models

Using LRT for this...

The \( Ldh \) gene family is an important model system for molecular evolution of isozyme multigene families. The rate of evolution is known to have increased in \( Ldh-C \) following the gene duplication event

1- an increase in the underlying mutation rate of \( Ldh-C \)
2- burst of positive selection for functional divergence following the duplication event
3- a long term change in selection pressure
Lactate dehydrogenase gene duplication?

And the LRT envelope please...

Parameter estimates under models of variable \( \alpha \) ratios among lineages and LRTs of their fit to the Ldh-A and Ldh-C gene family:

<table>
<thead>
<tr>
<th>Models</th>
<th>( \alpha_2 )</th>
<th>( \alpha_1 )</th>
<th>( \alpha_0 )</th>
<th>( \beta )</th>
<th>( F )</th>
<th>LRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>H(_0): ( \alpha_2 = \alpha_1 = \alpha_0 )</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>-601.63</td>
</tr>
<tr>
<td>H(_1): ( \alpha_2 = \alpha_1 \neq \alpha_0 )</td>
<td>0.13</td>
<td>0.13</td>
<td>0.14</td>
<td>0.13</td>
<td>0.19</td>
<td>-607.57</td>
</tr>
<tr>
<td>H(_2): ( \alpha_2 \neq \alpha_1 \neq \alpha_0 )</td>
<td>0.06</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>-598.11</td>
</tr>
<tr>
<td>H(_3): ( \alpha_2 \neq \alpha_1 = \alpha_0 )</td>
<td>0.06</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>-598.11</td>
</tr>
</tbody>
</table>

* The topology and branch specific \( \alpha \) ratios are presented in Figure 5.

1 H\(_0\) v H\(_1\), df = 1
2 H\(_0\) v H\(_2\), df = 1
3 H\(_0\) v H\(_3\), df = 1

H\(_2\) most likely...seems to be a ‘burst’
MBE toolbox in Matlab

http://www.pmarneffei.hku.hk/mbetoolbox/Download/login.asp

1. Alignment File IO

First, let's just show how to read a FASTA formatted file into a MATLAB structure to represent an alignment.

```matlab
Aln = readFasta('avg_examples\HOV_aligned.fas','2.1')
Reading avg_examples\HOV_aligned.fas
39
26
Aln = ...
genotype: 1
sequences: {1x21 '' '22''}
seq: {1x21} double
```

Now let's view the sequences in this alignment.

Notice MBEToolbox display sequences FMT/P format by default.

```matlab
viewSequences(Aln)
39
26
AAGACAGAGCTTCAAATCTCAAAAAAAAGAAAAAACATCACTCTGACGTCCGTCGAGCTG
AAGACAGAGCTTCAAATCTCAAAAAAAAGAAAAAACATCACTCTGACGTCCGTCGAGCTG
GAGTCAAGCTCAGGCGCTTCCAGTGACGTGGTGTTAGGTTCATCCTTCGCGG
GAGTCAAGCTCAGGCGCTTCCAGTGACGTGGTGTTAGGTTCATCCTTCGCGG
```

---

Given a number of aligned sequences of known length, both the nucleotide diversity and
the number of segregating sites can be calculated.

**Example:**

```
<table>
<thead>
<tr>
<th>Sequences (Total Length: 500)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
```

- Total segregating sites = 8
- One or more sequences must have a different nucleotide than the
  consensus sequence at a particular site for a segregating site to
  occur.
- Number of Alleles = 8
- Number of Alleles = 8
  
The number of alleles is determined by comparing entire sequences (as opposed to one nucleotide each). Sequences 4 and 5 represent the same allele.

- Frequency of Alleles (x)
  
  \[ x_1, x_2, \ldots, x_7 = 1/6 \times 0.11 \]
  
  \[ x_8, x_9 = 0.22 \]

- Calculate Pairwise Number of Segregating Sites (k)


<table>
<thead>
<tr>
<th>Pairs</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>

- Calculate Pairwise Nucleotide Differences \( d_{ij} \)


<table>
<thead>
<tr>
<th>Pairs</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>
Tajima's Test of Neutrality

This function conducts Tajima's test of neutrality (Tajima 1989). Tajima's test compares the number of segregating sites per site with the nucleotide diversity. (A site is considered a segregating site if there are two or more nucleotides at the site in a comparison of m sequences, and nucleotide diversity is defined as the average number of nucleotide differences per site between two sequences.) If all the alleles are selectively neutral, then the product \(4Nv\) (where \(N\) is the effective population size and \(v\) is the mutation rate per site) can be estimated in two ways, and the difference in the estimate obtained in these two ways provides indications of non-neutral evolution.

Nucleotide Diversity \( \frac{1}{m(m-1)} \sum \frac{x_i}{m} \) = \(0.68/0.047 = 0.0253\)

Let's read an example file:

```
Aln = readPHYLIP_T(1, "seq_examples/ing-new.phy", 2, 1)
```

Reading seq_examples/ing-new.phy
human-ECP
crane-ECP
gorilla-ECP
orang-ECP
maca-ECP
human-EDN
chimp-EDN
gorilla-EDN
orang-EDN
maca-EDN
mamba-EDN
```
Aln = seqtype: 2
    geneticcode: 1
    sequences: 1 (all seq)
    seqs: 11 x 671 (double)
```

Now, calculating Jukes & Cantor 69 distance. Matrix containing pairwise distances can be displayed.

```
D = dn_jukes_cantor69(Aln);
printMatrix(D, Aln)
```
Compute synonymous and non-synonymous substitution rates using Nei-Graur method

```plaintext
dB = dC = dNei_Graur(#Aln);  
printMatrix(dB, #Aln)
```

<table>
<thead>
<tr>
<th></th>
<th>human-ECF</th>
<th>chimp-ECF</th>
<th>gorila-ECF</th>
<th>orang-ECF</th>
<th>human-EKN</th>
<th>gorila-EKN</th>
<th>orang-EKN</th>
<th>chimp-EKN</th>
<th>macaq-EKN</th>
<th>tamar-EKN</th>
</tr>
</thead>
<tbody>
<tr>
<td>human-ECF</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>chimp-ECF</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td>gorila-ECF</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>orang-ECF</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>human-EKN</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>gorila-EKN</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>orang-EKN</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>chimp-EKN</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>macaq-EKN</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td>0.40</td>
<td>0.40</td>
<td>0.40</td>
<td>0.40</td>
<td>0.40</td>
<td>0.40</td>
</tr>
<tr>
<td>tamar-EKN</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
</tr>
</tbody>
</table>

113