**Distance estimation for nucleotide substitutions**

**Jukes-Cantor (one-parameter) method**

Jukes and Cantor (1969)

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

\[p(\lambda) = \frac{3}{4} \left( \frac{1 - p}{1 - 2p} \right) / L\]

All substitutions occur with equal probability

(Kules-Cantor model of nucleotide substitutions)

(Derivation of the JC equation: a note on Canvas)

**Kimura two-parameter method**

Kimura (1990)

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

\[p(\lambda) = \frac{3}{4} \left( \frac{1 - p}{1 - 2p} \right) / L\]

Difference in Ts and Tv substitutions (usually Ts > Tv) can be considered

(Kimura 2-parameter model of nucleotide substitutions)

**TODAY’S TOPICS**

- Distance estimation
  - for DNA sequences
  - for protein sequences
- Base composition bias, saturation
- Gamma distance
- Synonymous & nonsynonymous distances

**Assignment 9**
Sequence evolution as Markov process

Markov Chain: a discrete-time stochastic process

In more general continuous time scale,
- Markov Process

Sequence evolution as Markov process

Jukes-Cantor model of sequence evolution

Distance estimation for amino acid substitutions

PAM or Dayhoff distance

Distance estimation for amino acid substitutions

p: the proportion of amino acid differences when two sequences have n% distance
µ: the equilibrium frequency of the amino acid
**PAM matrices**

<table>
<thead>
<tr>
<th>Observed Distance</th>
<th>Evolutionary Distance</th>
<th>% actual distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>30</td>
<td>1.2</td>
</tr>
<tr>
<td>30</td>
<td>38</td>
<td>1.2</td>
</tr>
<tr>
<td>40</td>
<td>47</td>
<td>1.2</td>
</tr>
<tr>
<td>50</td>
<td>56</td>
<td>1.2</td>
</tr>
<tr>
<td>65</td>
<td>67</td>
<td>1.2</td>
</tr>
<tr>
<td>75</td>
<td>79</td>
<td>1.2</td>
</tr>
<tr>
<td>80</td>
<td>20</td>
<td>0.9</td>
</tr>
<tr>
<td>85</td>
<td>22</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Dayhoff et al. (1978)

**Distance estimation for amino acid substitutions**

- **Dayhoff**
- \(-\ln(p)\)
- \(-\ln(p-0.2)\)

(approximation of Dayhoff distance; Accurate when \(p < 0.75\))

Dayhoff matrices

**Distance estimation and assumptions**

- All nucleotide (or aa) sites change independently (Violation) e.g., Correlated changes within rRNA stem regions
- The substitution rate is constant over time and among different lineages
- The substitution rate is the same among all sites (Violation) e.g., Different DNA regions have different substitution rates
  - Use only sites with consistent substitution rates (synonymous vs. nonsynonymous; 1st, 2nd, 3rd codon positions)
  - Use distance methods that consider rate-heterogeneity among sites (Gamma distribution: e.g., Jin and Nei, Tamura and Nei methods, etc.)

**Distance estimation and assumptions**

- The base composition is at equilibrium
  - Among the sequences compared base composition is assumed to be the same
  - LogDet method is designed to circumvent this problem

**Choosing distance estimation methods**

- Which distance method should we choose?
  - Things to consider:
    - Base composition bias
    - Substitution pattern (Ts/Tv, etc.)
    - Rate-heterogeneity among sites

**Nucleotide substitution patterns: Mt vs. Nuclear**

**Table 1.** Base composition (%) of mtDNA and nuclear DNA genes of B. wolffii

<table>
<thead>
<tr>
<th>Gen</th>
<th>Length (bp)</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>Total</th>
<th>Synonymous sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyt</td>
<td>1318</td>
<td>416</td>
<td>471</td>
<td>364</td>
<td>373</td>
<td>1325</td>
<td>579</td>
</tr>
<tr>
<td>AT-rich</td>
<td>256</td>
<td>162</td>
<td>66</td>
<td>10</td>
<td>78</td>
<td>250</td>
<td>123</td>
</tr>
<tr>
<td>GC-rich</td>
<td>1056</td>
<td>464</td>
<td>94</td>
<td>74</td>
<td>417</td>
<td>1045</td>
<td>266</td>
</tr>
</tbody>
</table>

Nucleotide substitution patterns: Mt vs. Nuclear

Base composition is biased

Kimura 2-parameter method: no base composition bias


Choosing distance estimation methods

Which distance method should we choose?

- Things to consider:
  - Base composition bias
  - Substitution pattern (Ts/Tv, etc.)
  - Rate-heterogeneity among sites

Choosing distance estimation methods

Data:

Ts/Tv ≈ 2
0.3A:0.4T:0.2C:0.1G

Tamura distance:
Kimura 2-parameter + unequal GC%

Nucleotide substitution patterns: codon positions

Ts/Tv >> 2
Ts is saturated

Effect of multiple substitutions

Saturation plot

Evolve rapidly

Evolve slowly

Table 3.3 & 3.4
Observed numbers of nucleotide pairs and estimated distances between the human and Rhesus monkey mitochondrial cytochrome b genes. In Nei and Kumar (2000) "Molecular Evolution and Phylogenetics"

Table 3.5 & 3.6
Observed numbers of nucleotide pairs and estimated distances between the human and Rhesus monkey mitochondrial cytochrome b genes. In Nei and Kumar (2000) "Molecular Evolution and Phylogenetics"
Effect of multiple substitutions

Saturation plot

Genes or proteins with no saturation effect

Saturate quickly

Saturate slowly

Evolving without saturation

Transitions vs. transversions

Sequence data: Mitochondrial ND4 genes
Nematode species comparisons:
- a. within species
- b. within genera
- c. within families
- d. within orders


Transitions vs. transversions


Ts/Tv can be used as an indication of saturation

Choosing distance estimation methods

Which distance method should we choose?

Things to consider:
- Base composition bias
- Substitution pattern (Ts/Tv, etc.)
- Rate-heterogeneity among sites

Rate-heterogeneity among sites

Distance methods we discussed so far assume

the substitution rate is constant for all nucleotide or amino acid sites

In reality this assumption rarely holds:
- e.g., For protein-coding genes, 1st, 2nd, and 3rd codon positions (or synonymous vs. nonsynonymous sites) have different substitution rates.
  For RNA coding genes: loop vs. stem regions
  Functionally important vs. less important sites

Statistical analyses of the substitution rates suggest that the rate variation among different sites approximately follows a gamma distribution

Gamma distributions

When \( \alpha = \infty \):
- equal rate \( (r) \) for all sites
When \( \alpha = 1 \):
- \( r \) follows the exponential distribution
- \( r \) varies extremely
When \( \alpha < 1 \):
- many sites have \( r \) close to 0
- \( r \) practically invariable
Choosing distance estimation methods

- **Which distance method should we choose?**
  - Things to consider:
    - Base composition bias
    - Substitution pattern (Ts/Tv, etc.)
    - Rate-heterogeneity among sites
  - **Is including more parameters better?**
    - More flexible, more realistic
    - Larger sampling errors (lower statistical power)
    - More "undefined" distance problem
  
  e.g., if \( p \geq 0.75 \) in JC method \( k = -3/4 \ln(1-4p/3) \),
  \( k \) becomes "undefined" or "infinite"

Distance estimation and sampling error problem

<table>
<thead>
<tr>
<th>Uncorrected ( p )</th>
<th>JC distance</th>
<th>SE (100 bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.1073</td>
<td>0.03462</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2326</td>
<td>0.0545</td>
</tr>
<tr>
<td>0.3</td>
<td>0.3831</td>
<td>0.0764</td>
</tr>
<tr>
<td>0.4</td>
<td>0.5716</td>
<td>0.1050</td>
</tr>
<tr>
<td>0.5</td>
<td>0.8240</td>
<td>0.1500</td>
</tr>
<tr>
<td>0.6</td>
<td>1.2071</td>
<td>0.2449</td>
</tr>
<tr>
<td>0.66</td>
<td>1.5902</td>
<td>0.3948</td>
</tr>
<tr>
<td>0.7</td>
<td>2.0310</td>
<td>0.6874</td>
</tr>
<tr>
<td>0.72</td>
<td>2.4142</td>
<td>1.1225</td>
</tr>
<tr>
<td>0.74</td>
<td>3.2381</td>
<td>3.2989</td>
</tr>
</tbody>
</table>

Choosing distance estimation methods

- **jMODELTEST2** (for nucleotide substitution)
  - http://code.google.com/p/jmodeltest2/
  - A tool to carry out statistical selection of best-fit models of nucleotide substitution

- **PROTEOST3** (for amino acid substitution)
  - Abascal et al. (2005) Bioinformatics 21: 3196;
  - Baeurle et al. (2011) Bioinformatics 27: 1014
  - Amino acid substitution version of MODELLER

- **MEGA** http://www.megasoftware.net/
  - Model testing by Maximum Likelihood is available

Universal Genetic Code

- Synonymous (silent) substitutions DOES NOT change amino acids
- Nonsynonymous (replacement) substitutions DOES change amino acids
Synonymous/nonsynonymous distance methods

- Nei-Gojobori method (Nei and Gojobori, 1986)
  - Number of synonymous differences: $S_d$
  - Number of nonsynonymous differences: $N_d$
  - Proportion of synonymous differences: $p_S$
  - Proportion of nonsynonymous differences: $p_N$
  - $p_N = S_d/S_P$, $p_S = N_d/N$
  - $S$: Number of synonymous sites
  - $N$: Number of nonsynonymous sites
  - Jukes-Cantor correction for multiple-hits
    - $d_{JC} = -3/4ln(1-4p_N/3)$
    - $d_{JC} = -3/4ln(1-4p_S/3)$
  - K2P or Tajima-Nei (1 parameter-base freq.) correction is also used in modified versions

- How to count synonymous/nonsynonymous differences
  - One synonymous substitution
    - (Val) (Phe) (Leu)
    - (Val) (Val) (Arg)
  - Six possible pathways
  - Two possible pathways
  - 1/2 Nonsynonymous
  - 1/2 Synonymous
  - $d_{K2P}$ or $d_{Tajima-Nei}$ (1 parameter-base freq.) correction is also used in modified versions

Nucleotide substitution patterns

<table>
<thead>
<tr>
<th>Method</th>
<th>60-65%</th>
<th>65-70%</th>
<th>70-75%</th>
<th>75-80%</th>
<th>80-85%</th>
<th>85-90%</th>
<th>90-95%</th>
<th>95-100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC</td>
<td>0.40±0.06</td>
<td>0.89±0.04</td>
<td>1.61±0.00</td>
<td>5.01±0.02</td>
<td>5.90±0.05</td>
<td>7.87±0.06</td>
<td>8.92±0.06</td>
<td>10.54±0.08</td>
</tr>
<tr>
<td>LTR</td>
<td>0.16±0.04</td>
<td>0.89±0.04</td>
<td>1.61±0.00</td>
<td>5.01±0.02</td>
<td>5.90±0.05</td>
<td>7.87±0.06</td>
<td>8.92±0.06</td>
<td>10.54±0.08</td>
</tr>
<tr>
<td>New PBL-1</td>
<td>0.32±0.02</td>
<td>0.89±0.04</td>
<td>1.61±0.00</td>
<td>5.01±0.02</td>
<td>5.90±0.05</td>
<td>7.87±0.06</td>
<td>8.92±0.06</td>
<td>10.54±0.08</td>
</tr>
<tr>
<td>New PBL-2</td>
<td>0.32±0.02</td>
<td>0.89±0.04</td>
<td>1.61±0.00</td>
<td>5.01±0.02</td>
<td>5.90±0.05</td>
<td>7.87±0.06</td>
<td>8.92±0.06</td>
<td>10.54±0.08</td>
</tr>
<tr>
<td>Canoe</td>
<td>3.13±0.02</td>
<td>0.89±0.04</td>
<td>1.61±0.00</td>
<td>5.01±0.02</td>
<td>5.90±0.05</td>
<td>7.87±0.06</td>
<td>8.92±0.06</td>
<td>10.54±0.08</td>
</tr>
</tbody>
</table>

Available distance method programs

- ClustalW2 (ClustalX2)
  - K2P for DNA, hybrid sequence Kimura and PAM for protein
  - p < 0.05

- Nei-Gojobori method (Nei & Gojobori, 1986): based on JC model
  - LTR: Li-Wei-Lin method (Li et al., 1993): based on JC model
  - PBL or LTR: ParaBase (PBL) method (Parade and Baeuerle, 1993; Li, 1993)
  - K2P or Tajima-Nei (1 parameter-base freq.) correction is also used in modified versions

Phylogeny

- Available software and resources
  - ClustalW: http://www.clustal.org/clustalw2/
  - RAxML: http://www.cse.christiansen.dk/software/raxml/
  - MEGA: http://www.megasoftware.net/
  - Timeline: http://www.timelines.org/evolution/timeline.html
  - Phylogenetics: http://www.biology.gatech.edu/tutorials/phylogenetics/
  - YouTube tutorials available

- Phylogenetics on the Web
  - ClustalW2: http://www.ebi.ac.uk/clustalw2
  - RAxML: http://www.phylogeny.fr

- TreeBASE: http://www.treebase.org/

- MEGA 7: http://www.megasoftware.net/
  - Includes synonymous & nonsynonymous distances
  - PAML: http://abacusझू.ncr.ac.in/software/paml.html

- IP: web package for Analysis of Phylogenetics and Evolution
  - Includes many distance methods: http://ape-package.fr/
**Introduction to phylogeny**

➤ **Phylogeny (phylogenetic tree)**

- A graphic representation of evolutionary relationships among genes or organisms.
- True phylogeny cannot be known.
- We cannot actually observe the long-term evolution!
- Phylogenetic relationships can be only inferred.
- Phylogenetic relationships are reconstructed based on the information available (e.g., sequences).
- Represents a hypothesis of evolutionary relationships among gene or protein sequences: gene tree.
- Organismal relationships are inferred based on phylogenetic analysis: species tree.

**Note:** Gene trees do not always represent species trees!

**Tree terminology**

- **Terminal nodes (leaves)**
- Extant (existing) taxa: OTUs (operational taxonomic units).
- **Internal nodes**
- HTUs (hypothetical taxonomic units).

**Many ways of drawing trees**

- Only horizontal branches show the divergence level.

**Three different types of trees**

- **Cladogram**
  - Relative recency of common ancestry (or branching order).
  - No quantitative information.
- **Additive tree (phenogram)**
  - Branch lengths show the amount of evolutionary changes.
- **Ultrametric tree**
  - Branch length shows the amount of divergence.
  - In ultrametric trees, end nodes are all equidistant from the root of the tree.
  - Possible only assuming molecular clock (constant evolutionary rate).
  - Ultrametric tree: Branch length has no information.
  - Ultrametric tree: Branch length shows the amount of divergence.

**Many ways of drawing trees**

- Only vertical branches show the divergence level.
Introduction to phylogeny

Resolution of trees

- Star tree
  - No resolution
- Polytomy
- Partially resolved
- Fully resolved (bifurcating tree)

Nested parentheses format: Newick format

- Rooted
- Unrooted

Branch lengths

Unrooted tree

Rooted tree

Hypothetical Root

Root
Introduction to phylogeny

- **Outgroup**: used to "root" unrooted tree
  - Biological information is required to choose appropriate outgroup

Unrooted tree

Root for the ingroup

Rooted tree

Topography is the same!

- **Outgroup**: used to "root" unrooted tree
- Biological information is required to choose appropriate outgroup

Which species can be used as the outgroup?