Probabilistic modeling and molecular phylogeny

Anders Gorm Pedersen
Molecular Evolution Group
Center for Biological Sequence Analysis
Technical University of Denmark (DTU)
What is a model?

Mathematical models are:

- Incomprehensible
- Useless
- No fun at all
What is a model?

Mathematical models are:

- Incomprehensible
- Useless
- No fun at all
What is a model?

- Model = hypothesis !!!

- Hypothesis (as used in most biological research):
  - Precisely stated, but qualitative
  - Allows you to make qualitative predictions

- Arithmetic model:
  - Mathematically explicit (parameters)
  - Allows you to make quantitative predictions…
The Scientific Method

- Observation of data
- Model of how system works
- Prediction(s) about system behavior (simulation)
Modeling: An example
Modeling: An example

\[ y = ax + b \]

Simple 2-parameter model
Modeling: An example

\[ y = ax + b \]

Predictions based on model
Model Fit, parameter estimation

\[ y = ax + b \]

Measure of how well the model fits the data: sum of squared errors (SSE)

Best parameter estimates: those that give the smallest SSE (least squares model fitting)
Model Fit, parameter estimation

\[ y = 1.24x - 0.56 \]

Measure of fit between model and data: sum of squared errors (SSE)

Best parameter estimates: those that give the smallest SSE (least squares)
Probabilistic modeling applied to phylogeny

- Observed data: multiple alignment of sequences

  \[
  \begin{align*}
  \text{H. sapiens globin} & \quad \text{A} \ \text{G} \ \text{G} \ \text{G} \ \text{A} \ \text{T} \ \text{T} \ \text{C} \ \text{A} \\
  \text{M. musculus globin} & \quad \text{A} \ \text{C} \ \text{G} \ \text{G} \ \text{T} \ \text{T} \ \text{T} \ \text{-} \ \text{A} \\
  \text{R. rattus globin} & \quad \text{A} \ \text{C} \ \text{G} \ \text{G} \ \text{A} \ \text{T} \ \text{T} \ \text{-} \ \text{A}
  \end{align*}
  \]

- Probabilistic model parameters (simplest case):
  - Tree topology and branch lengths
  - Nucleotide-nucleotide substitution rates (or substitution probabilities)
  - Nucleotide frequencies: $\pi_A, \pi_C, \pi_G, \pi_T$

\[
Q = \begin{bmatrix}
  A & C & G & T \\
  A & -3\alpha & \alpha & \alpha & \alpha \\
  C & \alpha & -3\alpha & \alpha & \alpha \\
  G & \alpha & \alpha & -3\alpha & \alpha \\
  T & \alpha & \alpha & \alpha & -3\alpha
\end{bmatrix}
\]

\[
\Rightarrow P(t) = e^{Qt} = \begin{bmatrix}
P_A & P_A & P_A & P_A & P_A \\
P_A & P_A & P_A & P_A & P_A \\
P_A & P_A & P_A & P_A & P_A \\
P_A & P_A & P_A & P_A & P_A \\
P_A & P_A & P_A & P_A & P_A
\end{bmatrix}
\]

Probability matrix (function of time $t$)
The maximum likelihood approach I

- Starting point:
  You have some observed data and a probabilistic model for how the observed data was produced

- Example:
  - **Data**: result of tossing coin 10 times - 7 heads, 3 tails
  - **Model**: coin has probability $p$ for heads, $1-p$ for tails.
    The probability of observing $h$ heads among $n$ tosses is:
    \[
    P(h \text{ heads}) = \binom{n}{h} p^h (1 - p)^{n-h}
    \]

- Goal:
  You want to find the best estimate of the (unknown) parameter value based on the observations. (here the only parameter is “$p$”)
The maximum likelihood approach II

- Likelihood = Probability (Data | Model)

- Maximum likelihood:
  Best estimate is the set of parameter values which gives the highest possible likelihood.
Maximum likelihood: coin tossing example

Data: result of tossing coin 10 times - 7 heads, 3 tails

Model: coin has probability p for heads, 1-p for tails.

\[ P(\text{data}) = \binom{10}{7} p^7 (1 - p)^3 \]
Probabilistic modeling applied to phylogeny

• Observed data: multiple alignment of sequences

\[
\begin{align*}
\text{H. sapiens globin} & \quad AGGGATTCA \\
\text{M. musculus globin} & \quad ACGGTTTA \\
\text{R. rattus globin} & \quad ACGGATT-A
\end{align*}
\]

• Probabilistic model parameters (simplest case):
  – Tree topology and branch lengths
  – Nucleotide-nucleotide substitution rates (or substitution probabilities)
  – Nucleotide frequencies: \( \pi_A, \pi_C, \pi_G, \pi_T \)

\[
Q = \begin{bmatrix}
-3\alpha & \alpha & \alpha & \alpha \\
\alpha & -3\alpha & \alpha & \alpha \\
\alpha & \alpha & -3\alpha & \alpha \\
\alpha & \alpha & \alpha & -3\alpha
\end{bmatrix}
\]

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

Probability matrix (function of time \( t \))
Other models of nucleotide substitution

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Two-parameter model (Kimura 1980)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>1-α-2β</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>β</td>
<td>1-α-2β</td>
<td>β</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>β</td>
<td>α</td>
<td>1-α-2β</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>α</td>
<td>β</td>
<td>1-α-2β</td>
</tr>
<tr>
<td>2.</td>
<td>Four-parameter model (Blaisdell 1985)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>1-α-2γ</td>
<td>γ</td>
<td>α</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>δ</td>
<td>1-α-2δ</td>
<td>δ</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>δ</td>
<td>β</td>
<td>1-β-2δ</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>β</td>
<td>γ</td>
<td>1-β-2γ</td>
</tr>
<tr>
<td>3.</td>
<td>Six-parameter model (Kimura 1981a)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>1-2α-γ</td>
<td>γ</td>
<td>α</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>δ</td>
<td>1-2α-δ</td>
<td>δ</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>β</td>
<td>β</td>
<td>1-2β-ε</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>β</td>
<td>γ</td>
<td>1-2β-γ</td>
</tr>
<tr>
<td>4.</td>
<td>Nine-parameter model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>1-γβ₁-γα₁-γα₂</td>
<td>γβ₁</td>
<td>8Cγ₁</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>γβ₁</td>
<td>1-γβ₁-γα₁-γα₂</td>
<td>8Cα₂</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>γα₁</td>
<td>γβ₁-γα₁-γα₂-γα₃</td>
<td>8Cβ₂</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>γα₁</td>
<td>γβ₁-γα₁</td>
<td>1-γα₃-γα₂-γα₃</td>
</tr>
<tr>
<td>5.</td>
<td>General model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>1-α₁₂-α₂₃-α₄₃</td>
<td>α₁₃</td>
<td>α₄₃</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>α₂₃</td>
<td>1-α₁₂-α₂₃-α₄₃</td>
<td>α₂₄</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>α₂₄</td>
<td>1-α₁₃-α₂₃-α₄₃</td>
<td>α₃₄</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>α₃₄</td>
<td>α₄₃</td>
<td>1-α₄₃-α₁₃-α₄₃</td>
</tr>
</tbody>
</table>

*O, Original nucleotide; S, substitute nucleotide.
General Time Reversible Model

Table 13.3: The general time-reversible model of DNA evolution

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>$\pi_A \alpha$</td>
<td>$\pi_G \alpha$</td>
<td>$\pi_C \beta$</td>
<td>$\pi_T \gamma$</td>
</tr>
<tr>
<td>G</td>
<td>$\pi_A \alpha$</td>
<td></td>
<td>$\pi_C \delta$</td>
<td>$\pi_T \epsilon$</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>$\pi_A \beta$</td>
<td>$\pi_G \delta$</td>
<td></td>
<td>$\pi_T \eta$</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>$\pi_A \gamma$</td>
<td>$\pi_G \epsilon$</td>
<td>$\pi_C \eta$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Time-reversibility:

The amount of change from state $x$ to $y$ is equal to the amount of change from $y$ to $x$

$$\pi_A x P_{AG} = \pi_G x P_{GA}$$

$$\Rightarrow \pi_A x \pi_G x a = \pi_G x \pi_A x a$$
More elaborate models of evolution

- Codon-codon substitution rates
  (64 x 64 matrix of codon substitution rates)

- Different mutation rates at different sites in the gene
  (the “gamma-distribution” of mutation rates)

- Molecular clocks
  (same mutation rate on all branches of the tree).

- Different substitution matrices on different branches of the tree
  (e.g., strong selection on branch leading to specific group of animals)

- Etc., etc.
Different rates at different sites: the gamma distribution
Computing the probability of one column in an alignment given tree topology and other parameters

Columns in alignment contain homologous nucleotides

Assume tree topology, branch lengths, and other parameters are given. Assume ancestral states were A and A. Start computation at any internal or external node.

\[
Pr = \pi_c P_{CA}(t_1) P_{AC}(t_2) P_{AA}(t_3) P_{AG}(t_4) P_{AA}(t_5)
\]
Computing the probability of an entire alignment given tree topology and other parameters

- Probability must be summed over all possible combinations of ancestral nucleotides. (Here we have two internal nodes giving 16 possible combinations)

- Probability of individual columns are multiplied to give the overall probability of the alignment, i.e., the likelihood of the model.

- Often the log of the probability is used (log likelihood)
<table>
<thead>
<tr>
<th></th>
<th>Tree 1</th>
<th>Tree 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node 1</td>
<td>Node 2</td>
<td>Likelihood</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>0.00080905</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>0.00083558</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>0.00086109</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>0.00002653</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>0.00000034</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>0.00005659</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>0.00000555</td>
</tr>
<tr>
<td>C</td>
<td>T</td>
<td>0.0000017</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>0.00000046</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>0.00000719</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>0.00006010</td>
</tr>
<tr>
<td>G</td>
<td>T</td>
<td>0.00000023</td>
</tr>
<tr>
<td>T</td>
<td>A</td>
<td>0.00000011</td>
</tr>
<tr>
<td>T</td>
<td>C</td>
<td>0.00000180</td>
</tr>
<tr>
<td>T</td>
<td>G</td>
<td>0.00000185</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>0.00000168</td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td>0.00266831</td>
</tr>
<tr>
<td>Tree 1</td>
<td>Tree 2</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>Node 1</td>
<td>Node 2</td>
<td>Likelihood</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>0.00080905</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>0.00083558</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>0.00086109</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>0.00002653</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>0.00000034</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>0.00005659</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>0.00000555</td>
</tr>
<tr>
<td>C</td>
<td>T</td>
<td>0.00000017</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>0.00000046</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>0.00000719</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>0.00006010</td>
</tr>
<tr>
<td>G</td>
<td>T</td>
<td>0.00000023</td>
</tr>
<tr>
<td>T</td>
<td>A</td>
<td>0.00000011</td>
</tr>
<tr>
<td>T</td>
<td>C</td>
<td>0.00000180</td>
</tr>
<tr>
<td>T</td>
<td>G</td>
<td>0.00000185</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>0.00000168</td>
</tr>
<tr>
<td>Sum</td>
<td>0.00266831</td>
<td>Sum</td>
</tr>
</tbody>
</table>
Maximum likelihood phylogeny: heuristic tree search

- **Data:**
  - sequence alignment

- **Model parameters:**
  - nucleotide frequencies, nucleotide substitution rates, tree topology, branch lengths.

Choose random initial values for all parameters, compute likelihood

Change parameter values slightly in a direction so likelihood improves

Repeat until maximum found

**Results:**

1. ML estimate of tree topology
2. ML estimate of branch lengths
3. ML estimate of other model parameters
4. Measure of how well model fits data (likelihood).
Model Selection?

Measure of fit between model and data (e.g., SSE, likelihood, etc.)

How do we compare different types of models?

\[ y = 1.24x - 0.56 \]
Over-fitting: More parameters always result in a better fit to the data, but not necessarily in a better description.

\[ y = ax + b \]

2 parameter model
Good description, poor fit

\[ y = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g \]

7 parameter model
Poor description, good fit
Selecting the best model: the likelihood ratio test

- The fit of two alternative models can be compared using the ratio of their likelihoods:

\[ \text{LR} = \frac{P(\text{Data} \mid M_1)}{P(\text{Data} \mid M_2)} = \frac{L,M_1}{L,M_2} \]

- Note that LR > 1 if model 1 has the highest likelihood

- For nested models it can be shown that

\[ \Delta = 2\ln(\text{LR}) = 2\ln\left(L,M_1 - L,M_2\right) \]

follows a \( \chi^2 \) distribution with degrees of freedom equal to the number of extra parameters in the most complicated model.

This makes it possible to perform stringent statistical tests to determine which model (hypothesis) best describes the data.
Asking biological questions in a likelihood ratio testing framework

- Fit two alternative, nested models to the data.
- Record optimized likelihood and number of free parameters for each fitted model.
- Test if alternative (parameter-rich) model is *significantly* better than null-model, given number of additional parameters ($n_{\text{extra}}$):
  - Compute $\Delta = 2 \times (\ln L_{\text{Alternative}} - \ln L_{\text{Null}})$
  - Compare $\Delta$ to $\chi^2$ distribution with $n_{\text{extra}}$ degrees of freedom
- Depending on models compared, different biological questions can be addressed (presence of molecular clock, presence of positive selection, difference in mutation rates among sites or branches, etc.)
Likelihood ratio test example:
Which model fits best: JC or K2P?

<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>-</td>
<td>α</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>C</td>
<td>α</td>
<td>-</td>
<td>α</td>
<td>a</td>
</tr>
<tr>
<td>G</td>
<td>a</td>
<td>α</td>
<td>-</td>
<td>a</td>
</tr>
<tr>
<td>T</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>-</td>
</tr>
</tbody>
</table>

**Jukes and Cantor model (JC):**
- All nucleotides have same frequency
- All substitutions have same rate

<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>-</td>
<td>β</td>
<td>α</td>
<td>β</td>
</tr>
<tr>
<td>C</td>
<td>β</td>
<td>-</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td>G</td>
<td>α</td>
<td>β</td>
<td>-</td>
<td>β</td>
</tr>
<tr>
<td>T</td>
<td>β</td>
<td>α</td>
<td>β</td>
<td>-</td>
</tr>
</tbody>
</table>

**Kimura 2 parameter model (K2P):**
- All nucleotides have same frequency
- Transitions and transversions have different rate

=> K2P has one extra parameter compared to JC
Likelihood ratio test example:
Which model fits best: JC or K2P?

Starting point: set of DNA sequences, fit JC and K2P models to data, record likelihoods

<table>
<thead>
<tr>
<th>Model</th>
<th>lnL</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC</td>
<td>-2034.3</td>
</tr>
<tr>
<td>K2P</td>
<td>-2031.2</td>
</tr>
</tbody>
</table>

K2P has better fit than JC: \( \ln L_{K2P} > \ln L_{JC} \)

Test whether K2P is *significantly* better

\[
\Delta = 2 \times (\ln L_{\text{Alternative}} - \ln L_{\text{Null}}) = 2 \times (-2031.2 - -2034.3) = 6.2
\]

Degrees of freedom = 1
(K2P has one extra parameter compared to JC)
**Likelihood ratio test example:**

Which model fits best: JC or K2P?

\[
\Delta = 2 \times (\ln L_{\text{Alternative}} - \ln L_{\text{Null}}) = 6.2
\]

 Degrees of freedom = 1

<table>
<thead>
<tr>
<th>DF</th>
<th>0.9950</th>
<th>0.9750</th>
<th>0.9000</th>
<th>0.5000</th>
<th>0.1000</th>
<th>0.0500</th>
<th>0.0100</th>
<th>0.0010</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0000</td>
<td>0.0010</td>
<td>0.0158</td>
<td>0.4549</td>
<td>2.7055</td>
<td>3.8415</td>
<td>6.6349</td>
<td>10.8276</td>
</tr>
<tr>
<td>2</td>
<td>0.0100</td>
<td>0.0506</td>
<td>0.2107</td>
<td>1.3863</td>
<td>4.6052</td>
<td>5.9915</td>
<td>9.2103</td>
<td>13.8155</td>
</tr>
<tr>
<td>3</td>
<td>0.0717</td>
<td>0.2158</td>
<td>0.5844</td>
<td>2.3660</td>
<td>6.2514</td>
<td>7.8147</td>
<td>11.3449</td>
<td>16.2662</td>
</tr>
<tr>
<td>4</td>
<td>0.2070</td>
<td>0.4844</td>
<td>1.0636</td>
<td>3.3567</td>
<td>7.7994</td>
<td>9.4877</td>
<td>13.2767</td>
<td>18.4668</td>
</tr>
<tr>
<td>5</td>
<td>0.4117</td>
<td>0.8312</td>
<td>1.6103</td>
<td>4.3515</td>
<td>9.2364</td>
<td>11.0705</td>
<td>15.0863</td>
<td>20.5150</td>
</tr>
</tbody>
</table>

Critical value (5% level) = 3.8415

Statistic = 6.2

\[\Rightarrow 1\% < p < 5\%\]

\[\Rightarrow \text{Difference is significant}\]

\[\Rightarrow \text{K2P is significantly better description than JC}\]
Positive selection I: synonymous and non-synonymous mutations

- 20 amino acids, 61 codons
- Most amino acids encoded by more than one codon
  - Not all mutations lead to a change of the encoded amino acid
  - "Synonymous mutations" are rarely selected against

```
CTA (Leu)  
CTC (Leu)  
CTG (Leu)  
CTT (Leu)  
```

```
CGA (Arg)   CCA (Pro)   CAA (Gln)  
```

1 non-synonymous nucleotide site

1/3 synonymous
2/3 nonsynonymous nucleotide site

1/3 synonymous
2/3 nonsynonymous nucleotide site

1 synonymous nucleotide site
Positive selection II: non-synonymous and synonymous mutation rates contain information about selective pressure

- \(dN\): rate of non-synonymous mutations per non-synonymous site
- \(dS\): rate of synonymous mutations per synonymous site

Recall: Evolution is a two-step process:
(1) Mutation (random)
(2) Selection (non-random)

- Randomly occurring mutations will lead to \(dN/dS=1\).
- Significant deviations from this most likely caused by subsequent selection.

- \(dN/dS < 1\): Higher rate of synonymous mutations: negative (purifying) selection
- \(dN/dS > 1\): Higher rate of non-synonymous mutations: positive selection
Exercise: positive selection in HIV?

- Fit two alternative models to HIV data:
  - M1: two classes of sites with different dN/dS ratios:
    - Class 1 has dN/dS < 1
    - Class 2 has dN/dS = 1
  - M2: three distinct classes with different dN/dS ratios:
    - Class 1 has dN/dS < 1
    - Class 2 has dN/dS = 1
    - Class 3 has dN/dS > 1

- Use likelihood ratio test to examine if M2 is significantly better than M1.

- If M2 significantly better than M1 AND if some codons belong to the class with dN/dS > 1 (the positively selected class) then you have statistical evidence for positive selection.

- Most likely reason: immune escape (i.e., sites must be in epitopes)

○: Codons showing dN/dS > 1: likely epitopes