Models of Molecular Evolution

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

September 15, 2007
Features of Molecular Evolution

1. Possible multiple changes on edges
2. Transition/transversion bias
3. Non-uniform base composition
4. Rate variation across sites
5. Dependence among sites
6. Codon position
7. Protein structure
A Famous Quote About Models

Essentially, all models are wrong, but some are useful.

George Box
Probability Models

- A probabilistic framework provides a platform for formal statistical inference.
- Examining goodness of fit can lead to model refinement and a better understanding of the actual biological process.
- Model refinement is a continuing area of research.
- Most common models of molecular evolution treat sites as independent.
- These common models just need to describe the substitutions among four bases at a single site over time.
The Markov Property

- Use the notation $X(t)$ to represent the base at time $t$.
- Formal statement:

\[
P \{X(s + t) = j \mid X(s) = i, X(u) = x(u) \text{ for } u < s\} = P \{X(s + t) = j \mid X(s) = i\}
\]

- Informal understanding: given the present, the past is independent of the future
- If the expression does not depend on the time $s$, the Markov process is called \textit{homogeneous}.
Rate Matrix

- Positive off-diagonal rates of transition
- Negative total on the diagonal
- Row sums are zero
- Example

\[ Q = \{q_{ij}\} = \begin{pmatrix}
-1.1 & 0.3 & 0.6 & 0.2 \\
0.2 & -1.1 & 0.3 & 0.6 \\
0.4 & 0.3 & -0.9 & 0.2 \\
0.2 & 0.9 & 0.3 & -1.4 \\
\end{pmatrix} \]
Alarm Clock Description

- If the current state is $i$, the time to the next event is exponentially distributed with rate $-q_{ii}$ defined to be $q_i$.
- Given a transition occurs from state $i$, the probability that the transition is to state $j$ is proportional to $q_{ij}$, namely $q_{ij}/\sum_{k \neq i} q_{ik}$. 
Example: Begin at A, change to G at time 0.3, change to C at time 0.8, and then no more changes before time $t = 1$.

$$P\{\text{path}\} = P\{\text{begin at A}\} \\ \times \left(1.1e^{-(1.1)(0.3)} \cdot \frac{0.6}{1.1}\right) \\ \times \left(0.9e^{-(0.9)(0.5)} \cdot \frac{0.3}{0.9}\right) \\ \times \left(e^{-(1.1)(0.2)}\right)$$
For a continuous time Markov chain, the *transition matrix* whose \( ij \) element is the probability of being in state \( j \) at time \( t \) given the process begins in state \( i \) at time 0 is \( P(t) = e^{Qt} \).

A probability transition matrix has non-negative values and each row sums to one.

Each row contains the probabilities from a probability distribution on the possible states of the Markov process.
Examples

\[
P(0.1) = \begin{pmatrix}
0.897 & 0.029 & 0.055 & 0.019 \\
0.019 & 0.899 & 0.029 & 0.053 \\
0.037 & 0.029 & 0.916 & 0.019 \\
0.019 & 0.080 & 0.029 & 0.872
\end{pmatrix}
\]

\[
P(0.5) = \begin{pmatrix}
0.605 & 0.118 & 0.199 & 0.079 \\
0.079 & 0.629 & 0.118 & 0.174 \\
0.132 & 0.118 & 0.671 & 0.079 \\
0.079 & 0.261 & 0.118 & 0.542
\end{pmatrix}
\]

\[
P(1) = \begin{pmatrix}
0.407 & 0.190 & 0.276 & 0.126 \\
0.126 & 0.464 & 0.190 & 0.219 \\
0.184 & 0.190 & 0.500 & 0.126 \\
0.126 & 0.329 & 0.190 & 0.355
\end{pmatrix}
\]

\[
P(10) = \begin{pmatrix}
0.200 & 0.300 & 0.300 & 0.200 \\
0.200 & 0.300 & 0.300 & 0.200 \\
0.200 & 0.300 & 0.300 & 0.200 \\
0.200 & 0.300 & 0.300 & 0.200
\end{pmatrix}
\]
Well behaved continuous-time Markov chains have a stationary distribution, often designated $\pi$ (not the constant close to 3.14 related to circles).

When the time $t$ is large enough, the probability $P_{ij}(t)$ will be close to $\pi_j$ for each $i$. (See $P(10)$ from earlier.)

The stationary distribution can be thought of as a long-run average—over a long time, the proportion of time the state spends in state $i$ converges to $\pi_i$. 
The matrix $Q = \{q_{ij}\}$ is typically parameterized as $q_{ij} = r_{ij} \pi_j / \mu$ for $i \neq j$ which guarantees that $\pi$ will be the stationary distribution when $r_{ij} = r_{ji}$.
The expected number of substitutions per unit time is the average rate of substitution which is a weighted average of the rates for each state weighted by their stationary distribution.

\[ \mu = \sum_i \pi_i q_i \]

If the matrix \( Q \) is reparameterized so that all elements are divided by \( \mu \), then the unit of measurement becomes one substitution.
Time-reversibility

- The matrix $Q$ is the matrix for a time-reversible Markov chain when $\pi_i q_{ij} = \pi_j q_{ji}$ for all $i$ and $j$. That is the overall rate of substitutions from $i$ to $j$ equals the overall rate of substitutions from $j$ to $i$ for every pair of states $i$ and $j$. 