Outline

• Common nucleotide substitution models
• Ancestral sequence inference
Poisson process $\rightarrow$ continuous Markov process

$$\mathbb{P}[X_{t_{i+1}} = x_{i+1}|X_{t_i} = x_i] = e^{-\lambda(t_{i+1}-t_i)}(\lambda(t_{i+1} - t_i))^{x_{i+1}-x_i} \over (x_{i+1} - x_i)!$$

$$\mathbb{P}[Y_{t_{i+1}} = b|Y_{t_i} = a, X_{t_{i+1}} = x_{i+1}, X_{t_i} = x_i] = (R^{x_{i+1}-x_i})_{ab}$$
Jukes-Cantor Model

\[
\Pi = \frac{1}{4} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}
\]

\[
R = \frac{1}{4} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad Q = \frac{1}{4} \begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix}.
\]

\[
P(t) = \frac{1}{4} \begin{pmatrix} 1 + 3e^{-\lambda t} & 1 - e^{-\lambda t} & 1 - e^{-\lambda t} & 1 - e^{-\lambda t} \\ 1 - e^{-\lambda t} & 1 + 3e^{-\lambda t} & 1 - e^{-\lambda t} & 1 - e^{-\lambda t} \\ 1 - e^{-\lambda t} & 1 - e^{-\lambda t} & 1 + 3e^{-\lambda t} & 1 - e^{-\lambda t} \\ 1 - e^{-\lambda t} & 1 - e^{-\lambda t} & 1 - e^{-\lambda t} & 1 + 3e^{-\lambda t} \end{pmatrix}
\]
Properties of Jukes-Cantor

- **ergodic** \( \lim_{t \to \infty} P(t) = \Pi J \)
- **time reversible** \( \Pi Q = Q^T \Pi \)
Transitions vs. Transversions

- **Transition**
  - purine → purine
  - pyrimidine → pyrimidine

- **Transversion**
  - pyrimidine → purine
  - purine → pyrimidine
HKY85

\[
R = \begin{pmatrix}
1 - \pi_G \alpha - \pi_Y \beta & \pi_C \beta & \pi_G \alpha & \pi_T \beta \\
\pi_A \beta & 1 - \pi_G \alpha - \pi_R \beta & \pi_G \beta & \pi_T \alpha \\
\pi_A \alpha & \pi_C \beta & 1 - \pi_A \alpha - \pi_Y \beta & \pi_T \beta \\
\pi_A \beta & \pi_C \beta & \pi_G \beta & 1 - \pi_C \alpha - \pi_R \beta
\end{pmatrix},
\]

\[
Q = \begin{pmatrix}
-\pi_G \alpha - \pi_Y \beta & \pi_C \beta & \pi_G \alpha & \pi_T \beta \\
\pi_A \beta & -\pi_G \alpha - \pi_R \beta & \pi_G \beta & \pi_T \alpha \\
\pi_A \alpha & \pi_C \beta & -\pi_A \alpha - \pi_Y \beta & \pi_T \beta \\
\pi_A \beta & \pi_C \beta & \pi_G \beta & -\pi_C \alpha - \pi_R \beta
\end{pmatrix}.
\]

Here \(\alpha, \beta \geq 0\) and \(0 \leq \alpha + 2\beta < 1\) and \(\pi_A + \pi_C + \pi_G + \pi_T = 1\) with \(\pi_R = \pi_A + \pi_G\) and \(\pi_Y = \pi_C + \pi_T\).
Branch lengths

- What do the lengths of tree edges mean?
- Expected number of mutations per site
- time (t) and rate (λ) are not identifiable, only their product is: λt

\[ P(\text{mutation}|\text{event}) = 1 - \sum_{x \in \Sigma} \pi_x R_{xx} = 1 - \text{trace}(\Pi R) = -\text{trace}(\Pi Q). \]

Expected number of mutations = \( \lambda t \cdot (-\text{trace}(\Pi Q)) \)
Estimating branch length

- Given empirical $P(t)$, can estimate branch length with the $\logdet$ transform

\[
\logdet(P(t)) = -\log(\det(P(t)))
\]

\[
\logdet(P(t)) = \lambda t \cdot (-\text{trace}(Q))
\]
Tree model

\[
p(x_r) = \pi_{x_r}
\]

\[
p(x_j | x_i) = P_{x_i, x_j}(t_{ij})
\]

\[
p(x) = p(x_r) \prod_{(i,j) \in E} p(x_j | x_i)
\]

\[
= \pi_{x_r} \prod_{(i,j) \in E} P_{x_i, x_j}(t_{ij})
\]

\[
data = x_E = (x_1, x_2, x_3, x_4, x_5)
\]

model = \((T, P(t), \Pi)\)

\[
\mathbb{L}(data | model) = p(x_E) = \sum_{x_6, x_7, x_8, x_9} p(x)
\]
Likelihood computation

\(L_k(a):\) Likelihood of subtree rooted by \(k\) with \(x_k = a\)

for each vertex \(k\) in a topological ordering:

- if \(k\) is a leaf:
  
  set \(L_k(a) = \begin{cases} 
  1 & \text{if } x_k = a \\
  0 & \text{otherwise}
  \end{cases} \)

- else:

  let \(i\) and \(j\) be the child vertices of \(k\)
  
  set \(L_k(a) = \sum_{b,c} P_{ab}(t_{ki}) L_i(b) P_{ac}(t_{kj}) L_j(c)\)

return \(L(data|model) = \sum_a \pi_a L_r(a)\), where \(r\) is the root vertex

“Felsenstein’s algorithm” = sum-product algorithm
Marginal reconstruction

most likely character at root given the leaf characters

compute $L_r(a)$ using the likelihood algorithm
return $\arg\max_a \pi_a L_r(a)$
Joint reconstruction
character at root for most likely assignment to all interior vertices given the leaf characters

$\nabla_k(a)$: Likelihood of most likely interior vertex assignments for subtree rooted by $k$, with $x_k = a$

for each vertex $k$ in a topological ordering:
if $k$ is a leaf:
  set $\nabla_k(a) = \begin{cases} 1 & \text{if } x_k = a \\ 0 & \text{otherwise} \end{cases}$
else:
  let $i$ and $j$ be the child vertices of $k$
  set $\nabla_k(a) = \max_{b,c} P_{ab}(t_{ki})\nabla_i(b)P_{ac}(t_{kj})\nabla_j(c)$
return $\arg\max_a \pi_a \nabla_r(a)$, where $r$ is the root vertex