Outline

• Evolution on trees
• Markov chains
• Poisson processes
• Substitution matrices
Modeling evolution on trees - goals

• Given sequence data, we would like to:
  • Compare different tree topologies
  • Predict evolutionary scenarios
    • Reconstruct ancestral sequences
    • Identify mutation events
  • Estimate mutation rates
Modeling evolution on trees - methods

• Statistical methods
  • Continuous time Markov model
  • Poisson processes
  • Maximum likelihood
  • Tree graphical models
Statistical goal

• Given \( X \), a set of homologous nucleotide characters, we wish to model the generation of these characters on a tree

\[
P(X | T)
\]
Form of the models

\[ \pi_A, \pi_C, \pi_G, \pi_T \]

Root distribution

Branch lengths

\[ \mathbb{P}(A|\ell_1, \ldots) \]

Probability of substitution along each edge
Simple case

$X_r$: character state at root
$X_l$: character state at leaf

\[
\mathbb{P}[X_r = x_r, X_l = x_l] = \mathbb{P}[X_r = x_r] \mathbb{P}[X_l = x_l | X_r = x_r] = \pi_{x_r} \mathbb{P}_{x_r,x_l}
\]

\[
\mathbb{P}[X_r, X_l] = \Pi \mathbb{P}
\]
\( P[X_n = x|X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_0 = x_0] = P[X_n = x|X_{n-1} = x_{n-1}] \)

\( x \in \Sigma: \) the state space

\( P_{x,y} := P[X_{i+1} = y|X_i = x] \)

\( P^{(n)}_{x,y} = P[X_{i+n} = y|X_i = x] \)
Continuous time Markov process

**Definition 3.13** A continuous time Markov process is a sequence of $\Sigma$-valued random variable $\{Y_t, t \geq 0\}$ satisfying:

$$\mathbb{P}[Y_{t+s} = j|Y_t = i, Y_u = x_u, u \in [0, t)] = \mathbb{P}[Y_{t+s} = j|Y_t = i]$$  \hspace{1cm} (3.9)

for all $i, j \in \Sigma$ and $s, t > 0$. 
Counting process

Definition 3.1 (Counting process) A counting process is a sequence of random variables \( \{X_t, t \in \mathbb{R}^+ \cup \{0\}\} \) which satisfies:

1. \( X_t \) takes on values in \( \mathbb{Z}^+ \cup \{0\} \) for all \( t \),
2. \( X(s) \leq X(t) \) for \( s < t \),
3. \( \mathbb{E}[X_t] < \infty \),
4. \( X(t) - \lim_{h \to 0^+} X(t - h) \) is either 0 or 1.
Poisson process (Def I)

Definition 3.2 (Definition 1) A Poisson process with rate $\lambda \in \mathbb{R}^+$ is a counting process $\{X_t, t \geq 0\}$ satisfying:

1. For all $t_0 = 0 < t_1 < \cdots < t_n$, the differences $X_{t_1} - X_{t_0}, \ldots, X_{t_n} - X_{t_{n-1}}$ are independent random variables.

2. For all $s, t, u \geq 0$, $X_{s+t} - X_s \sim X_{u+t} - X_u$.

3. $\mathbb{P}[X_h = 1] = \lambda h + o(h)$, $\mathbb{P}[X_h \geq 2] = o(h)$.

4. $\mathbb{P}[X_0 = 0] = 1$. 

Poisson distribution

\[ P[X = k] = \frac{e^{-\lambda} \lambda^k}{k!} \]

\( \lambda > 0 \)

\( k = 0, 1, 2, \ldots \)

Limiting version of binomial distribution with \( n \) large, \( p \) small, and \( np = \lambda \)
Taylor series for $e$

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \ldots = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

converges for all $x$
Poisson process (Def II)

**Definition 3.5 (Definition 2)** A Poisson process with rate $\lambda \in \mathbb{R}^+$ is a counting process $\{X_t, t \geq 0\}$ satisfying:

1. For all $t_0 = 0 < t_1 < \cdots < t_n$, the differences $X_{t_1} - X_{t_0}, \ldots, X_{t_n} - X_{t_{n-1}}$ are independent random variables.

2. For all $s, t \geq 0$, $X_{s+t} - X_s \sim \text{Poisson}(\lambda t)$.

3. $\mathbb{P}[X_0 = 0] = 1$. 
Poisson process (Def III)

Definition 3.6 (Definition 3) A Poisson process with rate $\lambda \in \mathbb{R}^+$ is a counting process $\{X_t, t \geq 0\}$ satisfying:

1. For all $t_0 = 0 < t_1 < \cdots < t_n$, the differences $X_{t_1} - X_{t_0}, \ldots, X_{t_n} - X_{t_{n-1}}$ are independent random variables.

2. For all $s, t \geq 0$, $\mathbb{P}[X_{s+t} - X_s = 0] = e^{-\lambda t}$.

3. $\mathbb{P}[X_0 = 0] = 1$. 
Poisson→Continuous time Markov process

- State transition at each Poisson event
- Transition according to an event substitution probability matrix, $R$

$X_t$: number of events by time $t$

$\mathbb{P}[X_t = k] = \frac{e^{-\lambda t} (\lambda t)^k}{k!}$

$R = \begin{pmatrix}
    r_{AA} & r_{AC} & r_{AG} & r_{AT} \\
    r_{CA} & r_{CC} & r_{CG} & r_{CT} \\
    r_{GA} & r_{GC} & r_{GG} & r_{GT} \\
    r_{TA} & r_{TC} & r_{TG} & r_{TT}
\end{pmatrix}$

$\sum_j r_{ij} = 1$
Matrix exponentiation

For a square matrix, $A$

$$e^A := \sum_{k=0}^{\infty} \frac{A^k}{k!}$$

Exercise 3.10 Show that

1. $e^{cI} = e^{cI}$ for any $c \in \mathbb{R}$.
2. $e^{A+B} = e^A e^B$ if $A$ and $B$ commute.
3. $\det(e^A) = e^{\text{trace}(A)}$.
4. If $A$ is invertible then $e^{ABA^{-1}} = Ae^B A^{-1}$.
Substitution matrix derivation

\[ P(t) = \sum_{k=0}^{\infty} R^k \mathbb{P}[X_t = k] \]

\[ = \sum_{k=0}^{\infty} R^k \frac{(\lambda t)^k e^{-\lambda t}}{k!} \]

\[ = e^{-\lambda t} \mathbb{I} \sum_{k=0}^{\infty} R^k \frac{(\lambda t)^k}{k!} \]

\[ = e^{-\lambda t} \sum_{k=0}^{\infty} R^k \frac{(\lambda t)^k}{k!} \]

\[ = e^{(-I) \lambda t} \sum_{k=0}^{\infty} \frac{(R \lambda t)^k}{k!} \]

\[ = e^{(-I) \lambda t} e^{R \lambda t} = e^{Q \lambda t} \quad Q = R - I, \]
Rate matrix: $Q$

The matrix $Q = R - I$ satisfies

$$q_{ij} \geq 0 \quad \text{for } i \neq j,$$

$$\sum_{j \in \Sigma} q_{ij} = 0 \quad \text{for all } i \in \Sigma,$$

$$q_{ii} < 0 \quad \text{for all } i \in \Sigma.$$
Substitution matrix: \( P \)

**Exercise 3.12** Let \( Q \) be any matrix satisfying the conditions of Lemma 3.11 and let \( P(t) = e^{Q \lambda t} \). Show that

1. \( P(s + t) = P(s) \cdot P(t) \) \hspace{1em} (Chapman–Kolmogorov equations),

2. \( P(t) \) is the unique solution to the forward differential equation
   \[ P'(t) = \lambda P(t) \cdot Q, \quad P(0) = I \text{ for } t \geq 0 \]

3. \( P(t) \) is the unique solution to the backward differential equation
   \[ P'(t) = \lambda Q \cdot P(t), \quad P(0) = I \text{ for } t \geq 0, \]

4. \( P^{(k)}(0) = Q^k \).

5. A matrix \( Q \) is a rate matrix if and only if the matrix \( P(t) = e^{\lambda Q t} \) is a stochastic matrix (non-negative with row sums equal to one) for every \( t \geq 0 \).
Substitution matrix \( \rightarrow \) Continuous time Markov process

\[
P_{ij}(t) = \mathbb{P}[Y_t = j | Y_0 = i]
\]

Probability that nucleotide has changed from character \( i \) to character \( j \) in time \( t \).

- Hides “silent” events of two types:
  - self transitions (e.g., \( \text{A} \rightarrow \text{A} \))
  - back mutation (e.g., \( \text{A} \rightarrow \text{C} \rightarrow \text{A} \))