Substitution Models

Tomáš Vinař & Luca Denti

November 6, 2025

Markov chains

- similar to hidden Markov models: states and transitions, but no emissions
- more formally, sequence of random variables $X_0, X_1, \ldots, X_n$ such that state at time t depends only on the state at time $t-1$ and not on previous states:

$$P(X_t|X_0, \ldots, X_{t-1}) = P(X_t|X_{t-1})$$

- **Homogeneous Markov chains**: $P(X_t|X_{t-1})$ does not depend on t
- **Transition probability matrix**: probabilities of moving from one state to another

$$M[x, y] = P(X_t = y|X_{t-1} = x)$$

Example: states {A, C, G, T} can be used to track mutations at a specific position in a chromosome at a specific time point

Markov chains - Stationary distribution (/equilibrium)

- fundamental concept that describes the long-term behavior of a chain
- A distribution π over the set of states is called stationary for a Markov chain with transition matrix M if for every y it holds that

$$\sum_x \pi(x)M[x,y] = \pi(y)$$

  (or in matrix notation $\pi M = \pi$)

- after many steps ($t \to \infty$), all rows of the matrix converge to the stationary distribution
- A Markov chain can:
  - start in a stationary distribution and therefore remain in it
  - approach the equilibrium as time passes *(speed depends on chain structure)*
- A chain always converges (regardless of the initial state) if it is **ergodic**:

  M_t *for some* $t > 0$ *has all entries nonzero*

Substitution models, notation

$P(b|a,t)$: probability that if we start with symbol a, after time t we will see symbol b

Substitution models, notation

$P(b|a,t)$: probability that if we start with symbol a, after time t we will see symbol b

Transition probability matrix:

$$S(t) = \begin{pmatrix} P(A|A,t) & P(C|A,t) & P(G|A,t) & P(T|A,t) \\ P(A|C,t) & P(C|C,t) & P(G|C,t) & P(T|C,t) \\ P(A|G,t) & P(C|G,t) & P(G|G,t) & P(T|G,t) \\ P(A|T,t) & P(C|T,t) & P(G|T,t) & P(T|T,t) \end{pmatrix}$$

Substitution models, basic properties

- $S(0) = I$
- $\lim_{t\to\infty} S(t) = \begin{pmatrix} \pi_A & \pi_C & \pi_G & \pi_T \\ \pi_A & \pi_C & \pi_G & \pi_T \\ \pi_A & \pi_C & \pi_G & \pi_T \\ \pi_A & \pi_C & \pi_G & \pi_T \end{pmatrix}$

  Distribution π is called stationary (equilibrium)
- $S(t_1 + t_2) = S(t_1)S(t_2)$ (multiplicativity)
- Jukes-Cantor model should also satisfy

$$S(t) = \begin{pmatrix} 1-3s(t) & s(t) & s(t) & s(t) \\ s(t) & 1-3s(t) & s(t) & s(t) \\ s(t) & s(t) & 1-3s(t) & s(t) \\ s(t) & s(t) & s(t) & 1-3s(t) \end{pmatrix}$$

$$S(t) = \begin{pmatrix} 1-3s(t) & s(t) & s(t) & s(t) \\ s(t) & 1-3s(t) & s(t) & s(t) \\ s(t) & s(t) & 1-3s(t) & s(t) \\ s(t) & s(t) & s(t) & 1-3s(t) \end{pmatrix}$$

$$S(\epsilon) = \begin{pmatrix} 1-3s(\epsilon) & s(\epsilon) & s(\epsilon) & s(\epsilon) \\ s(\epsilon) & 1-3s(\epsilon) & s(\epsilon) & s(\epsilon) \\ s(\epsilon) & s(\epsilon) & 1-3s(\epsilon) & s(\epsilon) \\ s(\epsilon) & s(\epsilon) & s(\epsilon) & 1-3s(\epsilon) \end{pmatrix}$$

Jukes-Cantor model

$$S(t) = \begin{pmatrix} (1+3e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 \\ (1-e^{-4\alpha t})/4 & (1+3e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 \\ (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1+3e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 \\ (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1+3e^{-4\alpha t})/4 \end{pmatrix}$$

Equilibrium?

$$S(t) = \begin{pmatrix} 1-3s(t) & s(t) & s(t) & s(t) \\ s(t) & 1-3s(t) & s(t) & s(t) \\ s(t) & s(t) & 1-3s(t) & s(t) \\ s(t) & s(t) & s(t) & 1-3s(t) \end{pmatrix}$$

$$S(2t) = S(t)^2 =$$

$$= \begin{pmatrix} 1-6s(t)+12s(t)^2 & 2s(t)-4s(t)^2 & 2s(t)-4s(t)^2 & 2s(t)-4s(t)^2 \\ 2s(t)-4s(t)^2 & 1-6s(t)+12s(t)^2 & 2s(t)-4s(t)^2 & 2s(t)-4s(t)^2 \\ 2s(t)-4s(t)^2 & 2s(t)-4s(t)^2 & 1-6s(t)+12s(t)^2 & 2s(t)-4s(t)^2 \\ 2s(t)-4s(t)^2 & 2s(t)-4s(t)^2 & 2s(t)-4s(t)^2 & 1-6s(t)+12s(t)^2 \end{pmatrix}$$

$$\approx \begin{pmatrix} 1-6s(t) & 2s(t) & 2s(t) & 2s(t) \\ 2s(t) & 1-6s(t) & 2s(t) & 2s(t) \\ 2s(t) & 2s(t) & 1-6s(t) & 2s(t) \\ 2s(t) & 2s(t) & 2s(t) & 1-6s(t) \end{pmatrix}$$

for $t \to 0$

Substitution rate matrix (matica rýchlostí, matica intenzít)

- Substitution rate matrix for Jukes-Cantor model:

$$R = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

- For very small t we have $S(t) \approx I + Rt$

- Rate α is the probablity of a change per unit of time for very small t, or derivative of $s(t)$ with respect to t at $t = 0$

- Solving the differential equation for the Jukes-Cantor model we get $s(t) = (1 - e^{-4\alpha t})/4$

$$S(t) = \begin{pmatrix} 1-3s(t) & s(t) & s(t) & s(t) \\ s(t) & 1-3s(t) & s(t) & s(t) \\ s(t) & s(t) & 1-3s(t) & s(t) \\ s(t) & s(t) & s(t) & 1-3s(t) \end{pmatrix}$$

$$R = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

Jukes-Cantor model

$$S(t) = \begin{pmatrix} (1+3e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 \\ (1-e^{-4\alpha t})/4 & (1+3e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 \\ (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1+3e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 \\ (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1+3e^{-4\alpha t})/4 \end{pmatrix}$$

- The rate matrix is typically normalized so that there is on average one substitution per unit of time (time step = substitution occurred)
- In Jukes-Cantor model, $\alpha = 1/3$: it represents the equal probability of transition to any of the three other nucleotides from the current nucleotide

Jukes-Cantor model, summary

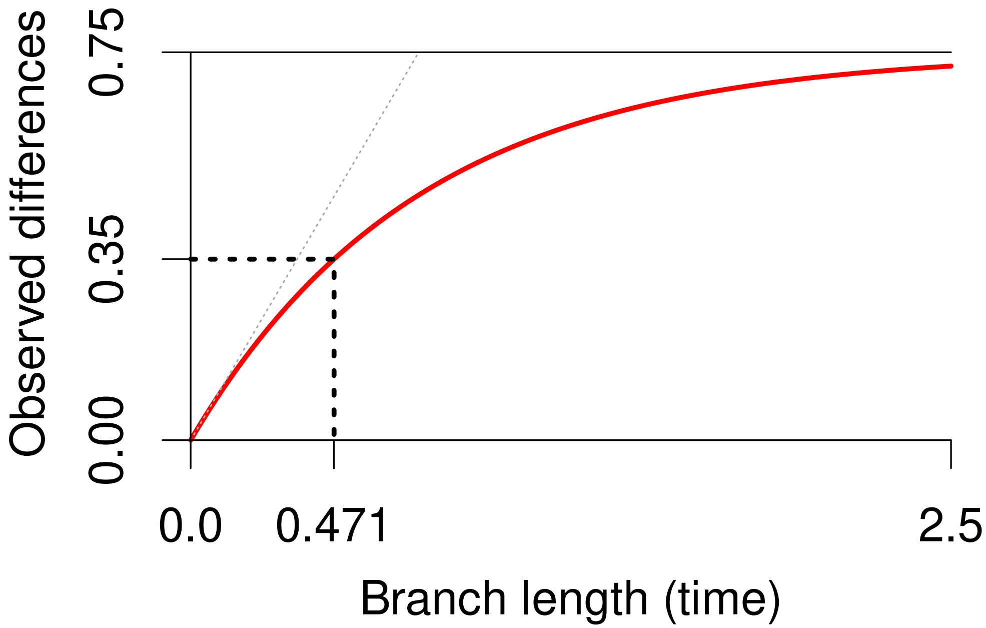
- $S(t)$: matrix 4×4, where $S(t)_{a,b} = P(b|a,t)$ is the probability that if we start with base a, after time t we have base b.
- Jukes-Cantor model assumes that $P(b|a,t)$ is the same for all $a \neq b$
- For a given time t, off-diagonal elements are $s(t)$, diagonal $1 - 3s(t)$
- Rate matrix R: for J-C off-diagonal α, diagonal -3α
- For very small t we have $S(t) \approx I - Rt$
- Rate α is the probablity of a change per unit of time for very small t, or derivative of $s(t)$ with respect to t for $t = 0$
- Solving the differential equation for the Jukes-Cantor model, we get $s(t) = (1 - e^{-4\alpha t})/4$
- The rate matrix is typically normalized so that there is on average one substitution per unit of time, that is, $\alpha = 1/3$

Correction of evolutionary distances

$$\Pr(X_{t_0+t} = C \mid X_{t_0} = A) = \tfrac{1}{4}(1 - e^{-\frac{4}{3}t})$$

The expected number of observed changes per base in time t:

$$D(t) = \Pr(X_{t_0+t} \neq X_{t_0}) = \tfrac{3}{4}(1 - e^{-\frac{4}{3}t})$$



Correction of observed distances

$$D = \frac{3}{4}\left(1 - e^{-\frac{4}{3}t}\right) \qquad \Rightarrow \qquad t = -\frac{3}{4}\ln\left(1 - \frac{4}{3}D\right)$$

More complex models

- General rate matrix R

$$R = \begin{pmatrix} \cdot & \mu_{AC} & \mu_{AG} & \mu_{AT} \\ \mu_{CA} & \cdot & \mu_{CG} & \mu_{CT} \\ \mu_{GA} & \mu_{GC} & \cdot & \mu_{GT} \\ \mu_{TA} & \mu_{TC} & \mu_{TG} & \cdot \end{pmatrix}$$

- μ_{xy} is the rate at which base x changes to a different base y
- Namely, $\mu_{xy} = \lim_{t \to 0} \frac{\Pr(y \mid x, t)}{t}$
- The diagonal is added so that the sum of each row is 0
- **General case:** 12 parameters. **Jukes-Cantor:** 1 parameter (α)
- There are models with a smaller number of parameters (compromise between Jukes-Cantor and an arbitrary matrix)

Kimura model

- A and G are purines, C and T pyrimidines
- Purines more often change to other purines

  and pyrimidines to pyrimidines
- Transition: change within group $A \Leftrightarrow G$, $C \Leftrightarrow T$,

  Transversion: change to a different group $\{A, G\} \Leftrightarrow \{C, T\}$
- Two parameters: rate of transitions α, rate of transversions β
- $R = \begin{pmatrix} -2\beta-\alpha & \beta & \alpha & \beta \\ \beta & -2\beta-\alpha & \beta & \alpha \\ \alpha & \beta & -2\beta-\alpha & \beta \\ \beta & \alpha & \beta & -2\beta-\alpha \end{pmatrix}$

HKY model (Hasegawa, Kishino, Yano)

- Extension of Kimura model, which allows different probabilities of A, C, G, T in the equilibrium
- If we set time to infinity, original base is not important, base frequencies stabilize in an equilibrium.
- Jukes-Cantor has probability of each base in the equilibrium 1/4.
- In HKY the equilibrium frequencies $\pi_A, \pi_C, \pi_G, \pi_T$ are parameters (summing to 1)
- Parameter κ: transition / transversion ratio (α/β)
- Rate matrix: $\mu_{x,y} = \begin{cases} \kappa\pi_y & \text{if mutation from } x \text{ to } y \text{ is transition} \\ \pi_y & \text{if mutation from } x \text{ to } y \text{ is transversion} \end{cases}$

From rate matrix R to transition probabilities $S(t)$

- J-C and some other models have explicit formulas for $S(t)$
- For more complex models, such formulas are not available
- In general, $S(t) = e^{Rt}$
- Exponential of a matrix A is defined as $e^{A} = \sum_{k=0}^{\infty} \frac{1}{k!} A^{k}$
- If R is diagonalized $R = UDU^{-1}$, where
  - D is a diagonal matrix with eigenvalues
  - U is the matrix of eigenvectors
- then $e^{Rt} = Ue^{Dt}U^{-1}$ and the exponential function is applied to the diagnal elements of D (more efficient)
- Note: diagonalization always exists for symmetric matrices R