Lecture 16

Gidon Rosalki

2025-12-14

1 Markov models

We can model nucleotide transitions using a Markov Chain. We will define the transition matrix of transition probabilities at the time t:

$$[\mathbb{P}[t]] = \begin{bmatrix} \mathbb{P}_t[A \to A] & \mathbb{P}_t[A \to C] & \mathbb{P}_t[A \to G] & \mathbb{P}_t[A \to T] \\ \mathbb{P}_t[C \to A] & \mathbb{P}_t[C \to C] & \mathbb{P}_t[C \to G] & \mathbb{P}_t[C \to T] \\ \mathbb{P}_t[G \to A] & \mathbb{P}_t[G \to C] & \mathbb{P}_t[G \to G] & \mathbb{P}_t[G \to T] \\ \mathbb{P}_t[T \to A] & \mathbb{P}_t[T \to C] & \mathbb{P}_t[T \to G] & \mathbb{P}_t[T \to T] \end{bmatrix} = \begin{bmatrix} \mathbb{P}_t & A & C & G & T \\ A & & & & \\ C & & & & \\ G & & & & \\ T & & & & \end{bmatrix} \tag{1}$$

$$\forall_{a,b} \mathbb{P}_t[a \to b] \geq 0 \tag{2}$$

$$\forall_a \sum_b \mathbb{P}_t(a \to b) = 1 \tag{3}$$

Additionally, from the Markov property:

$$\mathbb{P}[a \to b] = \sum_c (\mathbb{P}[a \to c] \cdot \mathbb{P}[c \to b])$$

Here we transition from a to c at time t_1, and from c to b at time t_2. We will also note that

$$\mathbb{P}[0] = I \tag{4}$$

$$\mathbb{P}[t_1 + t_2] = \mathbb{P}[t_1] \cdot \mathbb{P}[t_2] \tag{5}$$

$$\implies \mathbb{P}[2t_1] = \mathbb{P}[t_1]^2 \tag{6}$$

$$\implies \mathbb{P}[nt_1] = \mathbb{P}[t_1]^n \tag{7}$$

So given $\varepsilon > 0$, for all t, we may compute n such that $t = n\varepsilon$:

$$\mathbb{P}[t] = \mathbb{P}[\varepsilon]^{\frac{t}{\varepsilon}} = \mathbb{P}[\varepsilon]^n$$

For every time t, we can compute the transition matrix $\mathbb{P}[t]$, which is to say, the probability that some letter will transition to another letter. All that is left to do is evaluate the matrix. This leaves us with 3 problems:

1. How?
2. Exponentiating matrices is expensive
3. We have found a solution for $\mathbb{N}$, not $\mathbb{R}$

Firstly, recall how to differentiate from infi:

$$\begin{aligned} \frac{d\mathbb{P}[t]}{dt} &= \lim_{h \to 0} \frac{\mathbb{P}[t+h] - \mathbb{P}[t]}{h} \\ &= \lim_{h \to 0} \frac{\mathbb{P}[t] \cdot \mathbb{P}[h] - \mathbb{P}[t]}{h} \\ &= \mathbb{P}[t] \lim_{h \to 0} \frac{\mathbb{P}[h] - I}{h} \\ &= \mathbb{P}[t] \lim_{h \to 0} \frac{\mathbb{P}[0+h] - \mathbb{P}[0]}{h} \\ &= \mathbb{P}[t] \frac{d\mathbb{P}[0]}{dt} \end{aligned}$$

Which is to say, we have got a simple differential equation:

$$\frac{d\mathbb{P}[t]}{dt} = \mathbb{P}[t] \cdot \frac{d\mathbb{P}[0]}{dt} = \mathbb{P}[t] R$$

Where R is a rate matrix.

So,

$$\frac{d\mathbb{P}[t]}{dt} = \mathbb{P}[t] R \tag{8}$$

$$\mathbb{P}[0] = I \tag{9}$$

Which looks remarkably like how one differentiates the exponential:

$$f(0) = 1 \tag{10}$$

$$f'(x) = af(x) \tag{11}$$

Where f is e^{ax}. So, if we can compute $R = \frac{d\mathbb{P}[0]}{dt}$, then we can compute $\mathbb{P}[a \to b]$ at any time t.

We will note that this computation is $\mathbb{P}[t] = e^{tR}$. What does it mean to raise e to the power of a matrix? Well, 3blue1brown did a great `video` about this. Essentially, if we do Taylor series: $e^X = \sum_{n=0}^{\infty} \frac{X^n}{n!}$. Then we resolve the issue, since in our case it is simply that $X = (tR)$.

So, $R = \lim_{h \to 0} \frac{\mathbb{P}[h] - I}{h}$. Therefore, teh first approximation is $R \approx I + hR$. This gives us a first approximation where

$$R_{ij} = \begin{cases} \leq 0, & \text{if } i = j \\ > 0, & \text{if } i \neq j \end{cases} \tag{12}$$

$$\forall i \sum_j R_{i,j} = 0 \tag{13}$$

1.1 Jukes Cantor

In 1969, Jukes and Cantor created a simple approximation symmetric matrix:

$$R_{i,j} = \begin{cases} -3\alpha, & \text{if } i = j \\ \alpha, & \text{if } i \neq j \end{cases}$$

and the transition matrices:

$$\mathbb{P}[t]_{i,j} = \begin{cases} \frac{1}{4}\left(1 + 3e^{-4\alpha t}\right), & \text{if } i = j \\ \frac{1}{4}\left(1 - e^{-4\alpha t}\right), & \text{if } i \neq j \end{cases} \tag{14}$$

1.2 Kimura

In 1980, Kimura created something similar, but with differentiation between purines ($A \Leftrightarrow G$) and pyrimidine ($C \Leftrightarrow T$). We will note that transitions $\alpha >$ transversions β.

$$R = \begin{bmatrix} & A & C & G & T \\ A & -2\beta - \alpha & \beta & \alpha & \beta \\ C & \beta & -2\beta - \alpha & \beta & \alpha \\ G & \alpha & \beta & -2\beta - \alpha & \beta \\ T & \beta & \alpha & \beta & -2\beta - \alpha \end{bmatrix}$$

So the transition matrix at time t:

$$\mathbb{P}[t]_{i,j} = \begin{cases} 1 - 2s - u, & \text{if } i = j \\ u = \frac{1}{4}\left(1 + e^{-4\beta t} - 2e^{-2(\alpha+\beta)t}\right), & \text{if transition} \\ s = \frac{1}{4}\left(1 + e^{-4\beta t}\right), & \text{if transversion} \end{cases}$$

So, given a rate matrix, we may compute the probability of transition between every letter, for every possible time t. This is in other words, a simple model for evolution of sequences.