MATH 448 — Lecture 12
§4.4: Matrix models of base substitution

Terms

- **Transition matrix** M: the entry in row i, column j is $P(S_1 = i \mid S_0 = j)$, the probability that a site with base j in S_0 has base i in S_1. Bases are in the order A, G, C, T.
- **Markov matrix**: all entries are ≥ 0 and every column sums to 1.
- **Base distribution $\mathbf{p}_0$**: the column of probabilities that a site in S_0 has each base. Its entries are ≥ 0 and sum to 1.
- **Equilibrium $\mathbf{p}$**: a base distribution with $M\mathbf{p} = \mathbf{p}$, that is, an eigenvector with eigenvalue 1.

Rules

- One step: $\mathbf{p}_1 = M\mathbf{p}_0$. If the same M applies at every step, then $\mathbf{p}_2 = M\mathbf{p}_1$ and $\mathbf{p}_t = M^t\mathbf{p}_0$.
- A Markov matrix has eigenvalue 1, and every eigenvalue has $|\lambda| \leq 1$. If all entries are positive, 1 is the strictly dominant eigenvalue, there is one equilibrium, and $\mathbf{p}_t$ approaches it from any $\mathbf{p}_0$.
- **Jukes–Cantor model**: every substitution has probability $\alpha/3$, so the diagonal entries are $1 - \alpha$.

Problem 12.1. Use the sequences and count table of Problem 11.6. (a) Write the matrix M with entries $P(S_1 = i \mid S_0 = j)$, and the vector $\mathbf{p}_0$ of base probabilities in S_0. (b) Check that each column of M sums to 1 and that the entries of $\mathbf{p}_0$ sum to 1.

Solution. (a) Divide each column of the count table by its column sum, and divide the column sums by 40:

$$M \approx \begin{pmatrix} .667 & .385 & 0 & 0 \\ .111 & .538 & 0 & .111 \\ .222 & .077 & .667 & .222 \\ 0 & 0 & .333 & .667 \end{pmatrix}, \qquad \mathbf{p}_0 = \frac{1}{40}\begin{pmatrix} 9 \\ 13 \\ 9 \\ 9 \end{pmatrix} = \begin{pmatrix} .225 \\ .325 \\ .225 \\ .225 \end{pmatrix}.$$

(b) Each column of M is the conditional probabilities of the four bases in S_1 given one base in S_0, which add to 1. The entries of $\mathbf{p}_0$ are the probabilities of the four bases in S_0, which add to 1.

Problem 12.2. Compute the first entry of $\mathbf{p}_1 = M\mathbf{p}_0$ by multiplying, using fractions from the count table. Then find all of $\mathbf{p}_1$ from the row sums of the count table divided by 40. Why do they agree?

Solution. The first entry is $\frac{6}{9}\cdot\frac{9}{40} + \frac{5}{13}\cdot\frac{13}{40} + 0 + 0 = \frac{6}{40} + \frac{5}{40} = \frac{11}{40} = .275$. Each product $P(S_1 = i \mid S_0 = j)P(S_0 = j)$ equals $P(S_1 = i \text{ and } S_0 = j) = \text{count}_{ij}/40$, and the sum over j is the row total over 40. The row sums are $11, 9, 11, 9$, so $\mathbf{p}_1 = (.275, .225, .275, .225)$. These are the probabilities of the four bases in S_1.

Now do Exercise 12.1 from Homework 7.

Problem 12.3. Assume the same M applies in the next step, and compute $\mathbf{p}_2 = M\mathbf{p}_1$ with a calculator. Check that the entries sum to 1.

Solution. With $\mathbf{p}_1 = (.275, .225, .275, .225)$,

$$\mathbf{p}_2 = M\mathbf{p}_1 \approx (.270,\ .177,\ .312,\ .242),$$

for example $p_{2,\text{A}} = \frac{6}{9}(.275) + \frac{5}{13}(.225) = .183 + .087 \approx .270$. The rounded entries add to 1.001; without rounding they add to 1.

Problem 12.4. The Jukes–Cantor matrix with $\alpha = .03$ has .97 on the diagonal and .01 in every other position. (a) Check that each column sums to 1. (b) Compute $M\mathbf{q}$ for $\mathbf{q} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. What does this say about $\mathbf{q}$? (c) Compute $\mathbf{p}_1$ for $\mathbf{p}_0 = (.2, .3, .4, .1)$. (d) By what factor does the distance from each entry to $\frac{1}{4}$ change from $\mathbf{p}_0$ to $\mathbf{p}_1$?

Solution. (a) $.97 + 3(.01) = 1$. (b) Each entry of $M\mathbf{q}$ is $.97(\frac{1}{4}) + 3(.01)(\frac{1}{4}) = .25$, so $M\mathbf{q} = \mathbf{q}$: $\mathbf{q}$ is an equilibrium, an eigenvector with eigenvalue 1. (c) Entry i of $\mathbf{p}_1$ is $.97\,p_i + .01(1 - p_i) = .01 + .96\,p_i$, so $\mathbf{p}_1 = (.202,\ .298,\ .394,\ .106)$. (d) $p_{1,i} - \frac{1}{4} = .96\,(p_{0,i} - \frac{1}{4})$ for each i. For example, $.2 - .25 = -.05$ became $.202 - .25 = -.048$. The distance to equilibrium is multiplied by $.96 = 1 - \frac{4}{3}\alpha$ at each step.

Problem 12.5. A site is a purine or a pyrimidine. Let

$$M = \begin{pmatrix} .98 & .01 \\ .02 & .99 \end{pmatrix},$$

with the order (purine, pyrimidine). (a) Find the equilibrium base distribution. (b) Find the eigenvalues of M. (c) Start with a purine, $\mathbf{p}_0 = (1, 0)$, and find $\mathbf{p}_1$ and $\mathbf{p}_2$. How does the distance from $\mathbf{p}_2$ to the equilibrium compare with the distance from $\mathbf{p}_0$?

Solution. (a) Let $\mathbf{p} = (x, y)$ with $x + y = 1$. Then $.98x + .01y = x$ gives $.01y = .02x$, so $y = 2x$, and $\mathbf{p} = (\frac{1}{3}, \frac{2}{3})$. (b) The eigenvalues are 1 and, since the trace is 1.97, the other is .97. (Check: the determinant is $.9702 - .0002 = .97$.) Both columns of M are positive, so 1 is the strictly dominant eigenvalue. (c) $\mathbf{p}_1 = (.98, .02)$ and $\mathbf{p}_2 = (.9606, .0394)$. The first entry's distance from $\frac{1}{3}$ is .667 at time 0, .647 at time 1, and .627 at time 2. Each step multiplies the distance to equilibrium by .97, the second eigenvalue, so after two steps it is $.97^2 \approx .941$ times what it was.

Now do Exercise 12.2 from Homework 7.