HGEN 48600/STAT 35450: Fundamentals of Computational Biology

Lecture 6 — January 21

Lecturer: John Novembre *Scribe: Marcos Vieira*

Note: These lecture notes are still rough, and have only have been mildly proofread.

Discrete-time Markov Chains

$$X_1 \to X_2 \to ... \to X_t$$

Markov property:

$P(X_t = j \mid X_{t-1} = i_{t-1},\ X_{t-2} = i_{t-2},\ ...,\ X_1 = i_1) = P(X_t = j \mid X_{t-1} = i_{t-1})$

The chain is time-homogeneous if, for all t:

$P(X_t = j \mid X_{t-1} = i) = P_{ij}$

Since the system must move to one of the states:

$\sum_{j \in S} P_{ij} = 1$

We can collect the transition probabilities into a matrix $\mathbf{P}$. Because the rows of such a matrix specify probability distributions, the matrix is said to be a "stochastic matrix". For example:

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix}$$

An example from phylogenetics gives transition probabilities among nucleotides A, T, C and G, represented in the rows and columns:

$$\mathbf{P} = \begin{bmatrix} 0.999 & \frac{0.001}{3} & \frac{0.001}{3} & \frac{0.001}{3} \\ \frac{0.001}{3} & 0.999 & \frac{0.001}{3} & \frac{0.001}{3} \\ \frac{0.001}{3} & \frac{0.001}{3} & 0.999 & \frac{0.001}{3} \\ \frac{0.001}{3} & \frac{0.001}{3} & \frac{0.001}{3} & 0.999 \end{bmatrix}$$

An example from population genetics is the Wright-Fisher model for how the the number X_t of copies of an allele changes in a population over time:

$X_t \mid X_{t-1} = i \sim Binomial(N, \frac{i}{N})$

What happens in n steps?

$P_{ij}^{(n)}$: nth step transition probabilities.

$P_{ij}^{(n)} = P(X_{n+k} = j \mid X_k = i)$, for $n \geq 0$, $i, j \geq 0$

Chapman-Kolmogorov equations:

$P_{ij}^{(n+m)} = \sum_k P_{ik}{}^{(n)} P_{kj}{}^{(m)}$ for all $n, m \geq 0$

Or, equivalently, in matrix algebra:

$\mathbf{P}^{(n+m)} = \mathbf{P}^{(n)}\mathbf{P}^{(m)}$

In general, for discrete-time Markov chains:

$\mathbf{P}^{(n)} = \mathbf{P}^n$

Reducible vs. irreducible chains:

State j is *accessible* to state i if it's possible for the chain to move from i to j. If i is accessible to j, j is accessible to i, or both, states i and j *communicate*.

Class of states: a set of states that communicate.

Irreducible Markov chain: a Markov chain with a single class (all states communicate).

Recurrent or transient states:

f_i : probability of returning to i if starting at i, as $t \to \infty$

State i is:

Recurrent, if $f_i = 1$.
Transient, if $f_i < 1$.

Periodic vs. non-periodic states

A state is periodic if the chain can't stay in it and has to leave before moving to the same state again:

$P(X_{t+1} = i \mid X_t = i) = 0$

Otherwise, the state is non-periodic.

Ergodic Markov chains:

A discrete-time Markov chain is *ergodic* if it is irreducible and all of its states are recurrent and non-periodic.

If a discrete-time Markov chain is ergodic, then it's guaranteed to have a *stationary distribution* (also known as an equilibrium distribution). That is, from any initial probability distribution of the states $\boldsymbol{\pi}^{(0)}$, there is a unique $\boldsymbol{\pi}$ such that:

$$\lim_{t\to\infty} (\boldsymbol{\pi}^{(0)})^T \mathbf{P}^t = \boldsymbol{\pi}^T$$

And $\boldsymbol{\pi}$ satisfies:

$\boldsymbol{\pi}^T \mathbf{P} = \boldsymbol{\pi}$

Which we can solve for $\boldsymbol{\pi}$.

Eigenvalue decomposition of P:

If $\mathbf{V}$ and $\boldsymbol{\Lambda}$ are the eigenvector and eigenvalue matrices of $\mathbf{P}$, respectively, then:

$\mathbf{P} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$

$\mathbf{P}^k = \mathbf{V}\boldsymbol{\Lambda}^k\mathbf{V}^{-1}$

Usually, $\lambda_1 = 1$, $\lambda_2, \lambda_3, ... < 1$.

Time reversibility:

$$Q_{ij} = P(X_{t-1} = i \mid X_t = j) = \frac{P(X_t = i \mid X_{t-1} = j)P(X_{t-1} = j)}{P(X_t = i)}$$

If we are at the stationary distribution $\boldsymbol{\pi}$, this becomes:

$$Q_{ij} = \frac{P_{ij}\pi_j}{\pi_i}$$

In the special case where

$$P_{ij} = \frac{P_{ji}\pi_j}{\pi_i}$$

The chain is said to be *time-reversible*:

$\pi_i P_{ij} = \pi_j P_{ji}$