Markov Chains - Fall 2024

Instructor: Dr. Juhee Lee
TA: Antonio Aguirre

University of California, Santa Cruz
Fall 2024

Introduction

Markov chains are a versatile mathematical tool for modeling dynamic systems where transitions between states occur probabilistically. These models have applications in a wide range of fields, from biology to physics, and even traffic systems. In this document, we focus on one specific biological application: modeling DNA evolution.

DNA evolution involves random mutations, where a nucleotide at a given site in the genome (A, T, C, G) may change over time. Markov chains provide a structured way to analyze these changes, predict long-term nucleotide compositions, and explore equilibrium behavior. This guide introduces the concepts of transition matrices and stationary distributions, using DNA evolution as an intuitive example.

The Transition Matrix

A transition matrix P describes the probabilities of moving from one state to another in a single step. For DNA evolution, each state corresponds to a nucleotide (A, T, C, G). For example:

$$P = \begin{bmatrix} 0.85 & 0.05 & 0.05 & 0.05 \\ 0.10 & 0.80 & 0.05 & 0.05 \\ 0.10 & 0.05 & 0.80 & 0.05 \\ 0.10 & 0.05 & 0.05 & 0.80 \end{bmatrix}.$$

How to Read the Matrix

- Each row represents the **current nucleotide**. - Each column represents the **next nucleotide**. - For instance, $P_{AA} = 0.85$ indicates an 85% chance that A remains A in the next step, while $P_{AT} = 0.05$ means a 5% chance that A mutates to T.

Each row sums to 1 because the system must transition to some state ($\sum_j P_{ij} = 1$).

What Happens Over Time?

Initially, the nucleotide at a site might be A. After several transitions, it might mutate to T, C, or G. Over time, the system settles into a predictable pattern, regardless of its starting state. This is described by the **stationary distribution**.

The Stationary Distribution

The **stationary distribution** $\pi = [\pi_A, \pi_T, \pi_C, \pi_G]$ represents the long-term probabilities of each nucleotide appearing at the site. It satisfies:

$$\pi = \pi P, \quad \text{and} \quad \pi_A + \pi_T + \pi_C + \pi_G = 1.$$

Finding the Stationary Distribution

To compute π, solve:

$$\begin{aligned}
\pi_A &= 0.85\pi_A + 0.10\pi_T + 0.10\pi_C + 0.10\pi_G, \\
\pi_T &= 0.05\pi_A + 0.80\pi_T + 0.05\pi_C + 0.05\pi_G, \\
\pi_C &= 0.05\pi_A + 0.05\pi_T + 0.80\pi_C + 0.05\pi_G, \\
\pi_G &= 0.05\pi_A + 0.05\pi_T + 0.05\pi_C + 0.80\pi_G, \\
\pi_A + \pi_T + \pi_C + \pi_G &= 1.
\end{aligned}$$

Solving this system yields:

$$\pi = [0.4, 0.2, 0.2, 0.2].$$

What Does This Mean?

The stationary distribution indicates:

- $\pi_A = 0.4$: Over the long term, A will appear 40% of the time.
- $\pi_T = \pi_C = \pi_G = 0.2$: Each of T, C, G will appear 20% of the time.

If you observe this site for 1,000 time steps, approximately 400 steps will involve A, and 200 each will involve T, C, and G.

Why Is This Useful?

The stationary distribution has several practical applications:

- **Predicting DNA Composition:** It helps estimate the long-term nucleotide composition of a genome (e.g., GC content).
- **Validating Models:** By comparing observed nucleotide frequencies to the stationary distribution, we can evaluate how well a transition matrix models DNA evolution.
- **Phylogenetics:** Stationary distributions are integral to modeling mutation dynamics and constructing evolutionary trees.

Markov Chains and Equilibrium

Markov chains also allow us to model and explore equilibrium in broader contexts. **Equilibrium** refers to a stable, predictable pattern that emerges in the long run, regardless of the starting state.

Applications Beyond Biology

Markov chains are widely used in other fields to model dynamic systems:

- **Weather Prediction:** Modeling transitions between sunny, cloudy, and rainy days, with the stationary distribution indicating the probability of each weather pattern over time.
- **Traffic Flow:** Analyzing vehicle movement between intersections, with equilibrium describing stabilized traffic patterns.
- **Chemical Reactions:** Simulating molecular transitions between energy states, with the stationary distribution representing equilibrium compositions.
- **Population Dynamics:** Predicting long-term population balances in ecological systems, such as predator-prey relationships.

Conditions for the Stationary Distribution

For a stationary distribution to exist:

1. **Irreducibility:** Every state must be reachable from every other state.
2. **Aperiodicity:** The transitions must not follow strict cycles.
3. **Positive Recurrence:** The system must return to each state in a finite number of steps.
4. **Stochastic Matrix:** The rows of P must sum to 1.

Markov chains are an essential tool for studying dynamic systems, allowing us to predict equilibrium behavior in systems ranging from DNA evolution to weather and traffic. By using transition matrices and stationary distributions, we gain valuable insights into long-term patterns, even in highly complex systems.

Fun Fact: Who Was Markov?

The Markov chain is named after Andrey Andreyevich Markov (1856–1922), a Russian mathematician who made significant contributions to probability theory. Markov is best known for developing the concept of stochastic processes that bear his name—Markov chains.

Markov's work focused on sequences of random events where the future depends only on the present state and not on the history of past events. This property, now called the **Markov property**, became a cornerstone of modern probability theory.

Markov's first application of this idea was unconventional: he used it to analyze patterns in Russian poetry. Specifically, he studied the frequency and arrangement of vowels and consonants in texts by Alexander Pushkin. This demonstrated how mathematics could be used to study linguistic patterns, a novel idea at the time.

Markov was also known for his strong opinions. He famously disagreed with many of his contemporaries, including the mathematician Pavel Nekrasov, over how probability should be applied to real-world systems. Markov believed in using probability for practical problems, while others focused on philosophical interpretations.

He was also an outspoken critic of the Russian academic system under Tsarist rule, advocating for intellectual freedom in mathematics and science.

So, when you analyze DNA evolution using Markov chains, you're applying the same mathematical principles that Markov first used to study poetry more than a century ago!