Markov Chains - DNA Evolution

Antonio Aguirre

University of California, Santa Cruz
Winter 2025

Introduction

Markov chains provide a powerful mathematical framework for modeling dynamic systems where transitions between states occur probabilistically. These models are widely used in biology, physics, and computational sciences.

In this document, we focus on one specific biological application: **modeling DNA evolution**. DNA sequences undergo random mutations over time, where each nucleotide (A, T, C, G) may change into another with a certain probability. Markov chains help us analyze these transitions, predict long-term nucleotide compositions, and explore equilibrium behavior.

1. The Transition Matrix

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

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

How to Interpret the Matrix

- Each row represents the **current nucleotide**. - Each column represents the **next nucleotide**. - The entry P_{ij} represents the probability of transitioning from state i to state j. - Each row sums to 1: $\sum_j P_{ij} = 1$, ensuring total probability conservation.

Example: The probability of remaining in state A is 0.85, while the probability of transitioning from A to T is 0.05.

2. 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 a given site.

Stationary Distribution Condition

$$\pi = \pi P, \quad \text{subject to} \quad \sum \pi_i = 1.$$

Finding π

Solving:

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

yields:

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

Interpretation

- In the long run, A appears 40% of the time, while T, C, G each appear 20% of the time. - This equilibrium composition provides insights into long-term DNA stability.

3. Why Is This Useful?

Stationary distributions are essential for:

- **Predicting DNA Composition:** Estimating long-term nucleotide frequencies (e.g., GC content).
- **Validating Models:** Comparing observed DNA sequences to theoretical expectations.
- **Phylogenetics:** Modeling mutation rates for evolutionary tree construction.

4. General Markov Chain Properties

For a Markov chain to have a unique stationary distribution, it must satisfy:

1. **Irreducibility:** Every state must be reachable from any other state.
2. **Aperiodicity:** The system must not follow strict cycles.
3. **Positive Recurrence:** The system must return to each state in a finite number of steps.

5. Applications Beyond Biology

Today, Markov chains are widely used across scientific disciplines:

- **Genetics:** Modeling DNA sequence evolution.
- **Finance:** Forecasting stock market movements.
- **Artificial Intelligence:** Powering speech recognition systems like Siri and Google Assistant.
- **Physics and Chemistry:** Simulating molecular transitions and thermodynamic states.
- **Weather Forecasting:** Predicting climate patterns using stochastic models.

Fun Fact: Who Was Markov?

Andrey Andreyevich Markov (1856–1922) was a Russian mathematician whose pioneering work in probability theory led to the development of **Markov chains**. His research laid the groundwork for modern stochastic processes, with profound applications in fields as diverse as **biology, finance, artificial intelligence, and linguistics**.

The Birth of the Markov Property

Markov's research focused on sequences of random events where the probability of the next event depends only on the **present state** and not on past occurrences. This principle, now known as the **Markov property**, remains fundamental to probability theory and real-world applications.

The Markov Property

A stochastic process satisfies the **Markov property** if the probability of transitioning to the next state depends only on the current state, not on the sequence of states that preceded it.

This property allows Markov chains to model diverse systems, from predicting weather patterns to analyzing genetic mutations.

Markov and Poetry: An Unlikely Application

One of Markov's most unconventional applications of probability theory was in **literature**. In 1913, he analyzed the structure of Alexander Pushkin's poem *Eugene Onegin*, investigating the statistical dependence between vowels and consonants.

- His goal was to determine whether letters in a text could be treated as **independent random variables** or if they exhibited statistical dependence.
- He counted letter occurrences and transitions, demonstrating that the structure of written text follows probabilistic rules.
- This marked one of the earliest applications of probability to linguistics, paving the way for **computational linguistics and natural language processing (NLP)**.

Today, the principles Markov explored in poetry analysis are used in **chatbots, machine translation, and text generation algorithms**.

A Controversial Mathematician

Markov was known not only for his mathematical brilliance but also for his **outspoken nature**. He frequently clashed with other mathematicians, including Pavel Nekrasov, who believed probability should only apply to **independent events**. Beyond academia, Markov

was a vocal critic of the Russian Tsarist regime. He publicly opposed restrictions on academic freedom and took a strong stance against discrimination, particularly the expulsion of Jewish students from Russian universities. His **commitment to intellectual integrity and fairness** often put him at odds with authorities.

Markov Chains and DNA Evolution

When you use Markov chains to model DNA mutations, you are applying the same mathematical principles that Markov first used to study poetry over a century ago.

Markov's legacy proves that probability theory is not just about abstract mathematics—it is a **powerful tool for understanding complex systems**, from **literature to life sciences**.