JC69 Model Simulation Report

This report summarizes the approach and results of simulating the JC69 model of nucleotide substitution. The JC69 model assumes an equal probability of substitution between all nucleotides and is parameterized by a single substitution rate, alpha. This experiment aimed to observe how a DNA sequence converges to the stationary distribution of nucleotides over time under the JC69 model.

Methodology

The simulation begins with an initial DNA sequence and iteratively applies substitution steps according to the JC69 model parameters: alpha, a substitution rate, and a time interval, d. The probability of no substitution (remaining the same nucleotide) is calculated as: $p = 1/4 + 3/4 * \exp(-4 * \text{alpha} * d / 3)$

At each step, the frequency of each nucleotide is calculated, and the difference from the stationary distribution: [0.25, 0.25, 0.25, 0.25] is computed using the Euclidean metric.

Results

Four simulations were performed:
Two using natural sequence of Human Alcohol Dehydrogenase and two using artificial sequence with imbalanced nucleotides frequencies. Moreover, one simulation in each pair was done with d=0.05 and K=500 and second with d' = d/2, K'=K*2
All scenarios were plotted to observe convergence patterns and compared to assess the effect of rescaling.

Analysis

The Figures 1-2 show the progression of the sequence toward the stationary distribution over simulation steps. First of all, I noticed that for the natural sequence difference from stationary distribution is quite low from the beginning and converges very fast (Fig. 1). It is simply because sequence of Human Alcohol Dehydrogenase has well balanced nucleotide frequencies: [0.29, 0.21, 0.25, 0.25].
To observe more evident converging to stationary state, I tried artificial sequence with highly imbalanced frequencies: [0.69, 0.05, 0.08, 0.17]. Results are illustrated on Fig. 2. In that case, Euclidean distance clearly converges asymptotically to zero.
When it comes to effect of rescaling parameters d and K to d' = d/2 and K' = K*2. Final results are similar and distances converges on comparable level, however with rescaled parameters distance with converge slower.

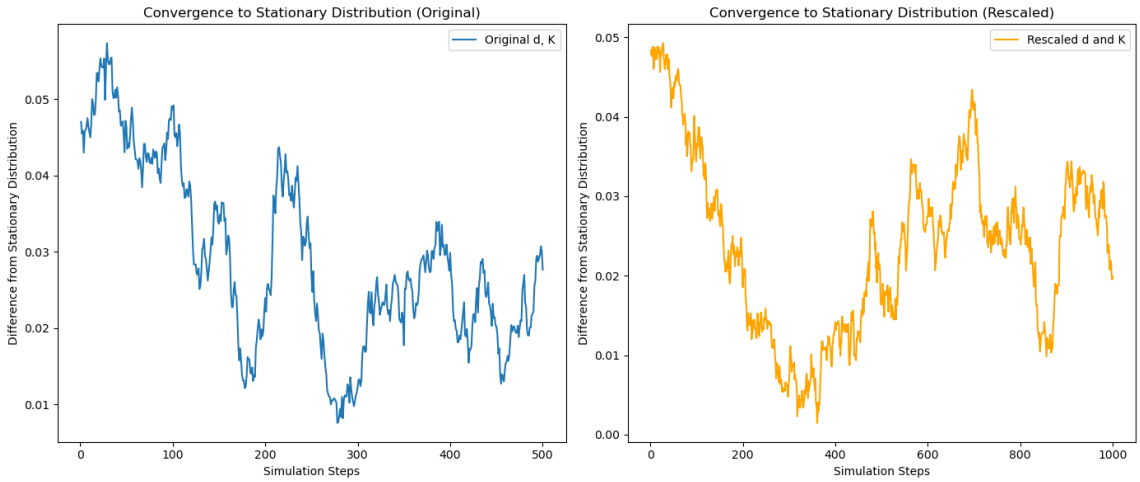

Figure 1. Visualization of Euclidean distance between stationary nucleotide frequencies and computed frequencies over K simulation steps for the natural sequence.

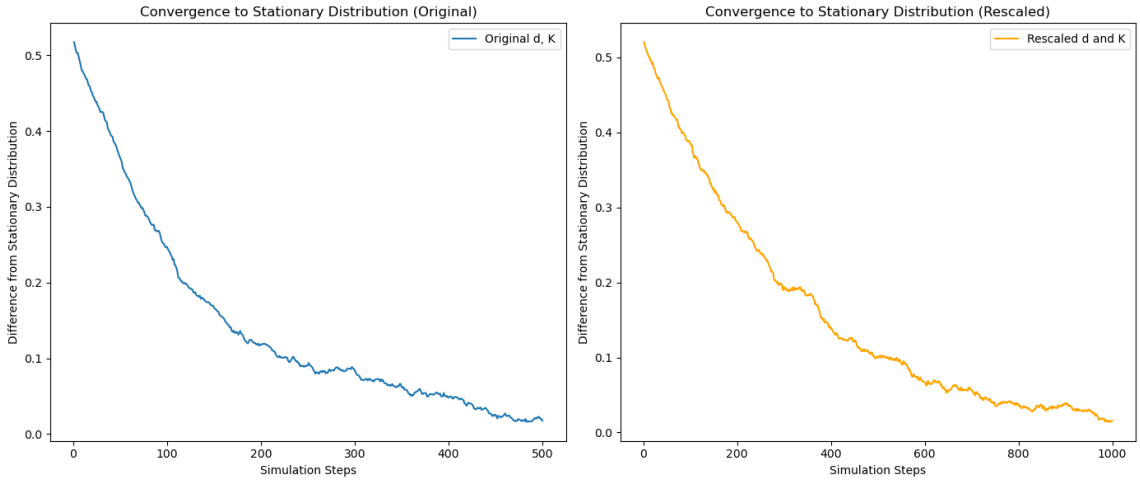

Figure 2. Visualization of Euclidean distance between stationary nucleotide frequencies and computed frequencies over K simulation steps for the artificial sequence.