Probabilistic Modeling in molecular evolution
A simulation study for the accuracy of parameters' estimation using K80 model

Damianos Melidis
dmelidis@student.ethz.ch

May 31, 2013

1. Abstract

In this task given that experimental sequences are too similar or too divergent we are trying to investigate the Kimura model's (K80) ability to estimate kappa by simulation of one sequences pair.

2. Design and Implementation of Simulation

2.a Designing

Firstly simulation design was governed by [Ziheng Yang] (chapters 1.1-1.2), consequently workflow has 6 (conceptual) steps. The following steps are performed for user specified number of replicates (n) and kappa (k_{user}), varying evolutionary distance d in the range [0.1,2.5] with step of 0.1. The first step generates the first sequence using equal probabilites for the four nucleotides ($p_A = p_G = p_T = p_C = 0.25$), then in the second step the probabilities of no change, transition and transversion (p_0, p_1 and p_2 respectively) is computed by the equations (1.10):

$$p_0 = \frac{1}{4} + \frac{1}{4} \cdot e^{\frac{-4d}{k+2}} + \frac{1}{2} \cdot e^{-2d\frac{k+1}{k+2}}$$
$$p_1 = \frac{1}{4} + \frac{1}{4} \cdot e^{\frac{-4d}{k+2}} - \frac{1}{2} \cdot e^{-2d\frac{k+1}{k+2}}$$
$$p_2 = \frac{1}{4} + \frac{1}{4} \cdot e^{\frac{-4d}{k+2}}$$

In the next step using the previous probabilities the second sequence is generated (evolved) by the first one independently for each base pair. Now having the sequences the empirical number of transition and transversion (denoted by S and V respectively) is calculated by counting the times of (A $\longleftrightarrow$ G or T $\longleftrightarrow$ C and A,G $\longleftrightarrow$ T,C). Afterwards using calculated S and V, assuming that 1 - 2S - V > 0, 1 - 2V > 0 and V $\neq 0$, kappa is estimated by the equation (1.11):

$$\hat{k} = \frac{2log(1-2S-V)}{log(1-2V)} - 1$$

Given the sequences' distance and user's kappa k_{user}, the simulation is performed 1000 times (n). For each simulation, the mean, the standard deviation and the mean square error of kappa estimation (of $\hat{k}$) is calculated by:

$$\bar{k} = \frac{1}{n}\sum_{i=1}^{n} \hat{k_i}$$
$$\sigma(\bar{k}) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (\hat{k_i} - \bar{k})^2}$$
$$mse(k) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{k_i} - k_{user})^2}$$

After all simulation of distances, the standard deviation of $\hat{k}$ versus the distance (d) is ploted, indicating also the user initial kappa.

2.b Implementing

In this section briefly the implementation scheme is described. Firstly the main (main.py) asks user for the kappa and maximum number of simulations per distance (the length of sequence is predefined to 1000 nucleotides). Then the initial sequence is generated (generate_initial_seq.py), followed by the creation of the second (evolve_second_seq.py). After having the two sequences the number of transition and transversion are counted and corresponding distance and kappa are estimated (count_transition_transversion.py). As the simulations for each distance end the mean, the standard deviation and the mean square error of kappa is calculated (compute_accuracy_statistics.py). In the final step a log file which contains for each simulation the accuracy statistics is created (Kimura_Simulation.txt) and the plot of estimated kappa versus the simulated distances (Kappa_est_over_dist.jpg) by plot_hist_k_hat.py.

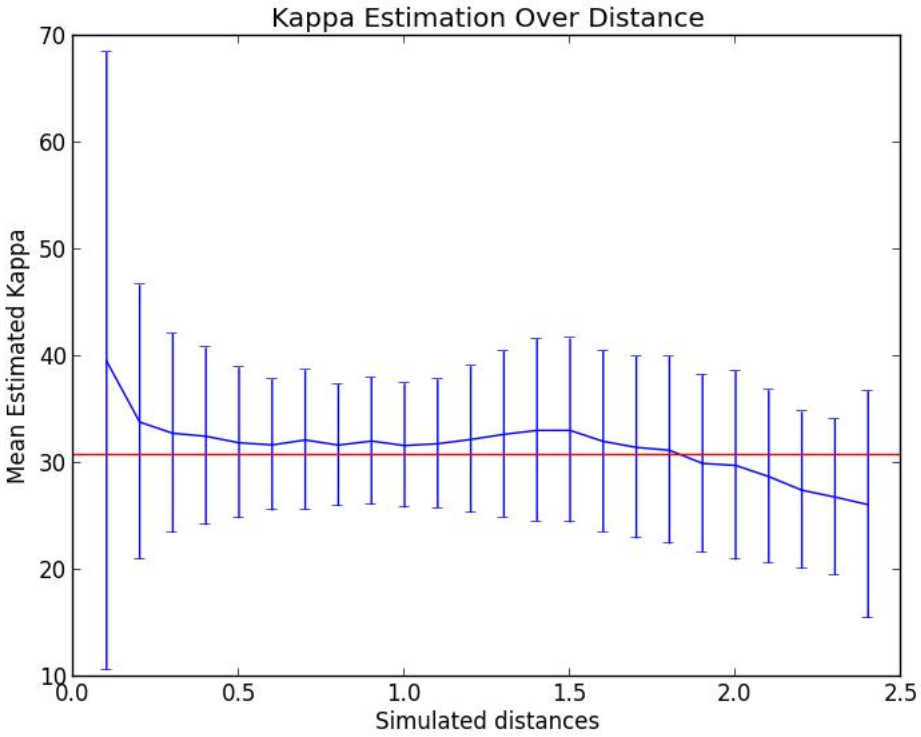

Figure 1: Standard deviation of $\hat{k}$ over distances (where red line corresponds to k_{user})

3. Results

The following results are drawn selecting $\kappa = 30.8$ (as this is the mean estimated kappa for Kimura model (table 1.3 of chapter 1.2) and the number of simulations per distance is equal to 1000. From the plot of $\hat{k}$ over distance (Figure 1.) we can observe that for distance in [0.5,1.1] the estimated κ is close to the input k_{user} and the standard deviation of the estimation is the lowest possible. In contrary, when the simulated distance is too low, namely [0.1, 0.3] or too high [2.1, 2.5] the model does estimate the κ with the maximum possible error and standard deviation. Consequently if our sequences are too similar or too divergent the estimation of transition-transversion ratio is uniformative.
It is important to report that increasing the distance, the estimation of kappa could be done less frequently, because 1 - 2S - V > 0, 1 - 2V > 0 and V $\neq 0$ do not hold. The number of simulations where the estimation of kappa is feasible is shown in the log file Kimura_Simulation.txt. Fortunately this phenomenon does not effect the comparison between $\hat{k}$ for different distances, as the number of simulations per distance is 1000.
In conclusion we can state that the biologist thought is verified by the previous experiment.

Bibliography

[Ziheng Yang] Ziheng Yang (2006) Computational Molecular Evolution