# Modeling transition/transversion bias of nucleotide substitution over time

Stefan Schmutz
January 2019

## Introduction

DNA, the molecule essential for known life, consists of four building blocks. Those blocks (thymine (T), cytosine (C), adenine (A), guanine (G)) are called nucleotides.[1]
Since DNA evolves over time, substitutions (change of nucleotides) occur. For structural reasons, transitions (substitutions between A and G or T and C), are more likely to happen compared to transversions (all other substitutions) in the majority of the cases[2].

The aim of this work is to describe a Markov model of nucleotide substitution over time which consideres nucleotide- and transition/transversion-bias.
The four nucleotides of DNA represent the four different states of the Markov model.
Other parameters needed for the model are the steady-state distribution ($\pi$) and state transition probabilities.
Those parameters will be estimated in the next section and are based on a pairwise sequence alignment of human and mouse cytochrome b (MT-CYB) which is a mitochondrial gene[3].

## Estimations to parameterize the model

There are several different Markov models of nucleotide substitution described in the literature[4]. The HKY85 model[5] was chosen because it fits the problem at hand.

### Nucleotide frequencies

> Estimate nucleotide frequencies from the pairwise alignment of human and mouse cytochrome b gene as given in the file "mt-cyb-human-mouse_cDNAalignment.fasta". Use these values to parameterize the model.

A way to estimate the nucleotide frequencies is to count the occurences and divide them by the total length. Since we're working with a pairwise alignment without indels, the total length of both sequences is the same (1140 nt). The detailed composition is listed in Table 1.

Table 1: Nucleotide frequencies

| nucleotide | human | mouse |
|---|---|---|
| T | 0.2509 | 0.2868 |
| C | 0.3430 | 0.2737 |
| A | 0.2860 | 0.3167 |
| G | 0.1202 | 0.1228 |

We assume that the mean of this distribution is the steady-state distribution $\pi$.

$$\pi = (\pi_T, \pi_C, \pi_A, \pi_G) = (0.2689, 0.3084, 0.3014, 0.1215)$$

**Transition transversion rate ratio**

> Propose a simple way of estimating transition transversion rate ratio from the dataset and use this estimate for the parameterization of the model.

The frequencies of the 16 possible combinations of the sequence alignment are shown in Table 2 (from mouse in rows to human in columns). We get the frequencies when the numbers (occurrences) are divided by the total length (1140 nt).

Table 2: Nucleotide comparisons (frequencies)

| | T | C | A | G |
|---|---|---|---|---|
| T | 0.2035 | 0.0632 | 0.0175 | 0.0026 |
| C | 0.0281 | 0.2184 | 0.0237 | 0.0035 |
| A | 0.0167 | 0.0518 | 0.2316 | 0.0167 |
| G | 0.0026 | 0.0096 | 0.0132 | 0.0974 |

The fraction of sites with transitional differences ($S = 0.1211$) and transversional differences ($V = 0.1281$) can be estimated by taking the sum of the corresponding fields from Table 2.
There are different definitions of the transition transversion rate ratio.[4] It was decided to apply the measure used by Kimura[6] and Hasegawa et al.[5] ($\kappa = \alpha/\beta$). The estimate of transition transversion rate ratio ($\hat{\kappa}$) can be calculated as follows[4]:

$$\hat{\kappa} = \frac{2\ln(1-2S-V)}{\ln(1-2V)} - 1 = 2.1248$$

**Substitution rate matrix**

The substitution rate matrix $Q$ of the HKY85 model[5] is defined as follows (rows and columns are ordered $T, C, A, G$):

$$Q = \{q_{ij}\} = \begin{bmatrix} . & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & . & \pi_A & \pi_G \\ \pi_T & \pi_C & . & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & . \end{bmatrix} \mu$$

The diagonal of $Q$ is defined by the requirement that each row of a substitution rate matrix must sum to 0. Using the steady-state distribution $\pi$ and transition transversion rate ratio $\kappa$ calculated above, we can fill in the rate matrix (without considering $\mu$ yet):

$$Q = \{q_{ij}\} = \begin{bmatrix} -1.078 & 0.655 & 0.301 & 0.121 \\ 0.571 & -0.994 & 0.301 & 0.121 \\ 0.269 & 0.308 & -0.835 & 0.258 \\ 0.269 & 0.308 & 0.640 & -1.217 \end{bmatrix} \mu$$

The factor $\mu$ rescales the matrix that the mean substitutionrate is one. It can be calculated as follows:[7]

$$\mu = \frac{-1}{\sum_{i \in \{T,C,A,G\}} \pi_i q_{ii}} = 1.004$$

$$Q = \{q_{ij}\} = \begin{bmatrix} -1.082 & 0.658 & 0.303 & 0.122 \\ 0.574 & -0.998 & 0.303 & 0.122 \\ 0.270 & 0.310 & -0.839 & 0.259 \\ 0.270 & 0.310 & 0.643 & -1.222 \end{bmatrix}$$

## (log-)likelihood

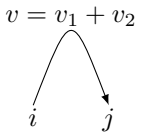
### Computing by hand

> By hand, compute the (log-)likelihood for the first 10 alignment positions given that the two sequences are separated by evolutionary distance of 0.01 expected substitutions per site.

To compute the likelihood ($L$) and log-likelihood ($\ell$) the transition probability matrix is computed from the substitution rate matrix $Q$ and evolutionary distance $v$ as follows:[7]

$$P(v) = e^{Qv} = \begin{bmatrix} 0.989 & 0.007 & 0.003 & 0.001 \\ 0.006 & 0.990 & 0.003 & 0.001 \\ 0.003 & 0.003 & 0.992 & 0.003 \\ 0.003 & 0.003 & 0.006 & 0.988 \end{bmatrix}$$

We're given that the distance of the two sequences from human and mouse is $v = 0.01$. This means that there are 0.01 expected substitutions between the two sequences per position.
Since we're only given one distance, we assume a unrooted tree where the mouse sequence is ancestral to the human sequence.




The likelihood is defined as the product of the probabilities for each site. To answer the question of the likelihood of the first ten alignment positions, we therefore need to compute these probabilities first:

$$Pr_{AA}(v, \kappa, \pi) = \pi_A * p_{AA}(v) = 0.301 * 0.992 = 0.299$$
$$Pr_{TT}(v, \kappa, \pi) = \pi_T * p_{TT}(v) = 0.269 * 0.989 = 0.266$$
$$Pr_{GG}(v, \kappa, \pi) = \pi_G * p_{GG}(v) = 0.121 * 0.988 = 0.120$$
$$Pr_{AA}(v, \kappa, \pi) = \pi_A * p_{AA}(v) = 0.301 * 0.992 = 0.299$$
$$Pr_{CC}(v, \kappa, \pi) = \pi_C * p_{CC}(v) = 0.308 * 0.990 = 0.305$$
$$Pr_{AC}(v, \kappa, \pi) = \pi_A * p_{AC}(v) = 0.301 * 0.003 = 0.001$$
$$Pr_{AC}(v, \kappa, \pi) = \pi_A * p_{AC}(v) = 0.301 * 0.003 = 0.001$$
$$Pr_{AC}(v, \kappa, \pi) = \pi_A * p_{AC}(v) = 0.301 * 0.003 = 0.001$$
$$Pr_{CA}(v, \kappa, \pi) = \pi_C * p_{CA}(v) = 0.308 * 0.003 = 0.001$$
$$Pr_{AA}(v, \kappa, \pi) = \pi_A * p_{AA}(v) = 0.301 * 0.992 = 0.299$$

$$L = \prod_{site=1}^{N} Pr_{site} = 1.976e - 16$$

To avoid underflow (computer issue caused by very small numbers) log-likelihood is often used:

$$\ell = \sum_{site=1}^{N} \ln(Pr_{site}) = -36.160$$

**Computing using a program**

> Write a program to compute the likelihood function of the whole alignment for the same genetic distance. Compute the likelihood for the whole alignment for several values of transition-transversion ratio around the value you estimated. Can you find a value that gives a better likelihood?

The function `log_likelihood` was written, which takes following arguments:

- `v`: evolutionary distance (expected substitutions per site)
- `kappa`: transition transversion rate ratio
- `pi`: vector of steady-state distribution of nucleotides (T, C, A, G)
- `df`: dataframe with alignment (organism in columns, sites in rows)

and returns the log-likelihood (using the estimated parameter values) of the sequence alignment which in this case is:

$$\ell = -3105.197$$

We can now fix all parameters except the transition transversion rate ratio $\kappa$ and look for the maximum log-likelihood (Figure 1).

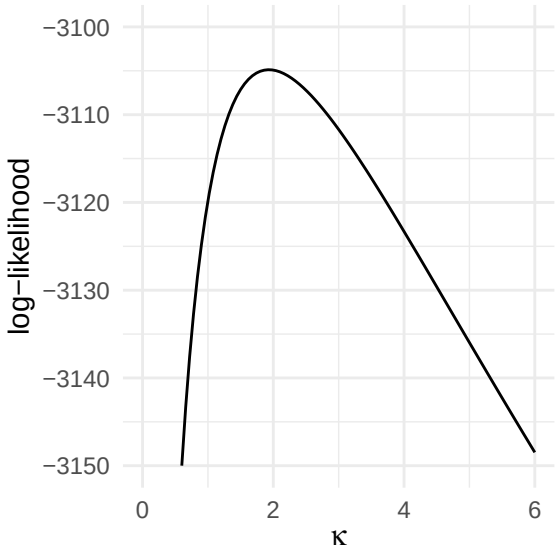



Figure 1: Log-likelihood curve of the proposed model

There's a transition transversion rate ratio which gives a slightly better likelihood:

$$\hat{\kappa} = 1.932$$
$$\ell = -3104.878$$

## Implementing a simulation

> Implement a simulation under your model for two sequences over an arbitrary distance $v$. Validate the program by simulation and show the results of you validation.

To simulate two aligned sequences over a distance of $v = 0.01$ and the previously defined substitution rate matrix $Q$ it's assumed that the substitutions happen independantly across sites.
1000 pairwise alignments of length 1140 nt will be simulated.

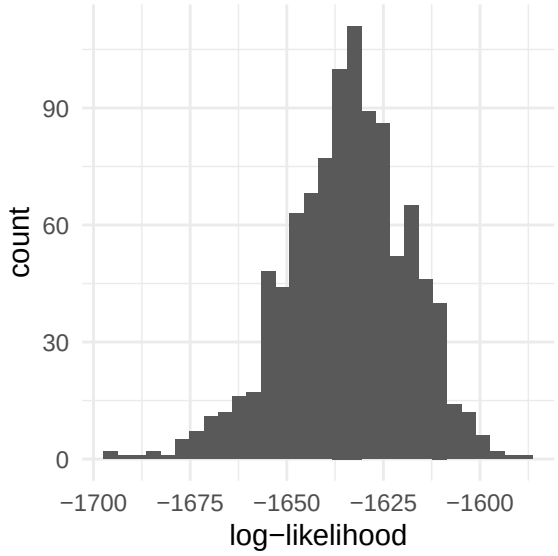



Figure 2: Log-likelihood distribution of simulated sequences

## Discussion

It's unclear if the (requested) applied model is a good choice for this example. The transition transversion bias of the given sequence alignment is not very strong. Furthermore more complex models usually require more data to infer the parameter and are more prone to overfitting.
The evolutionary distance given ($v = 0.01$) is much lower compared to the estimated distance $\hat{v}$ based on the K80 model[6]:

$$\hat{v} = -\frac{1}{2}\ln(1 - 2S - V) - \frac{1}{4}\ln(1 - 2V) = 0.3051$$

This therefore influences the probabilities of the transition matrix. Changing that distance results in a better likelihood as next section proves.

If we look for the optimum (maximum likelihood) of a combination of the two parameters $v$ and $\kappa$ we get following results:

$$\hat{v} = 0.306$$

$$\hat{\kappa} = 2.166$$

with a log-likelihood of $\ell = -2427.499$ which is higher than the previously calculated maximum likelihood.

## Disclosure

TODO: Mention the help I got. (Teja, optim with two parameter)

## Notation

| | |
|---|---|
| $\pi$ | steady-state distribution of nucleotides (frequencies) |
| $S$ | transitional differences (frequencies) |
| $V$ | transversional differences (frequencies) |
| $\kappa$ | transition transversion rate ratio |
| $\alpha$ | transition rate |
| $\beta$ | transversion rate |
| $Q$ | substitution rate matrix |
| $q_{ij}$ | substitution rate from nucleotide i to j |
| $q_{ii}$ | diagonal element of substitution rate matrix ($i = j$) |
| $\mu$ | rescaling factor |
| $P$ | transition probability matrix |
| $v$ | evolutionary distance (expected substitutions per site) |
| $L$ | likelihood |
| $\ell$ | log-likelihood ($\log_e$) |
| $Pr_{ij}$ | probability of observing nucleotide i and j |
| $p_{ij}$ | probability of substition of nucleotide i to j |

## References

TODO: cite R and packages


1. Wikipedia contributors. DNA. (2019).

2. Keller, D. A. N., Irene AND Bensasson. Transition-transversion bias is not universal: A counter example from grasshopper pseudogenes. *PLOS Genetics* **3**, 1–7 (2007).

3. Andreu, A. L. *et al.* Exercise intolerance due to mutations in the cytochrome b gene of mitochondrial dna. *New England Journal of Medicine* **341**, 1037–1044 (1999).

4. Yang, Z. Molecular evolution. A statistical approach. (2014) doi:10.1093/acprof:oso/9780199602605.001.0001.

5. Hasegawa, M., Kishino, H. & Yano, T.-a. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution* **22**, 160–174 (1985).

6. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**, 111–120 (1980).

7. John P., H. Calculating likelihoods on phylogenetic trees. (2010).