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) and 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 considers nucleotide- and transition/transversion-bias.
The four nucleotides of DNA represent the four states of the Markov model. Other parameters needed for the model are the steady-state distribution (π) and state transition probabilities ($\boldsymbol{P}$). 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 considers nucleotide- and transition/transversion-bias.

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 occurrences 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 nucleotide composition is listed in Table 1.

Table 1: Nucleotide frequencies

nucleotide	human	mouse	mean
T	0.251	0.287	0.269
C	0.343	0.274	0.308
A	0.286	0.317	0.301
G	0.120	0.123	0.121

We assume that the mean of this distribution is the steady-state distribution π. As $v \to \infty$ the equation $\pi \boldsymbol{P}(v) = \pi$ holds true[4].

$$\hat{\pi} = (\hat{\pi}_T, \hat{\pi}_C, \hat{\pi}_A, \hat{\pi}_G) = (0.269, 0.308, 0.301, 0.121)$$

Transition transversion rate ratio

> Propose a simple way of estimating transition transversion rate ratio from the data-set 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	0.204	0.063	0.018	0.003
C	0.028	0.218	0.024	0.004
A	0.017	0.052	0.232	0.017
G	0.003	0.010	0.013	0.097

The fraction of sites with transitional differences ($S = 0.121$) and transversional differences ($V = 0.128$) 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].
A simple way to estimate the transition transversion rate ratio (κ) is to divide the mean transitional differences by the mean transversional differences:

$$\hat{\kappa} = \frac{S/4}{V/8} = 1.890$$

If $\kappa = 1$ this means that there's no difference in transition and transversion substitutions (no bias). We see that our $\hat{\kappa}$ is slightly above 1 which indicates a bias towards transitions.

Substitution rate matrix

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

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

The diagonal of $\boldsymbol{Q}$ is defined by the requirement that each row of a rate matrix must sum to 0.
Using the steady-state distribution π and transition transversion rate ratio κ estimated above, we can fill in the rate matrix (without considering μ yet):

$$\boldsymbol{Q} = \{q_{ij}\} = \begin{bmatrix} -1.006 & 0.583 & 0.301 & 0.121 \\ 0.508 & -0.931 & 0.301 & 0.121 \\ 0.269 & 0.308 & -0.807 & 0.230 \\ 0.269 & 0.308 & 0.570 & -1.147 \end{bmatrix} \mu$$

The factor μ re-scales the matrix that the mean substitution rate is one. It can be calculated as follows:[6]

$$\mu = \frac{-1}{\sum_{i \in \{T,C,A,G\}} \pi_i q_{ii}} = 1.064$$

$$\boldsymbol{Q} = \{q_{ij}\} = \begin{bmatrix} -1.070 & 0.620 & 0.321 & 0.129 \\ 0.541 & -0.991 & 0.321 & 0.129 \\ 0.286 & 0.328 & -0.858 & 0.244 \\ 0.286 & 0.328 & 0.606 & -1.220 \end{bmatrix}$$

(log-)likelihood

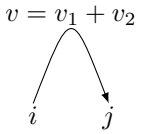
Computing by hand

> 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 (ℓ) the transition probability matrix is computed from the substitution rate matrix $\boldsymbol{Q}$ and evolutionary distance v as follows:[6]

$$\boldsymbol{P}(v) = e^{\boldsymbol{Q}v} = \begin{bmatrix} 0.989 & 0.006 & 0.003 & 0.001 \\ 0.005 & 0.990 & 0.003 & 0.001 \\ 0.003 & 0.003 & 0.991 & 0.002 \\ 0.003 & 0.003 & 0.006 & 0.988 \end{bmatrix}$$

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 an unrooted tree where the mouse sequence (i) is ancestral to the human sequence (j).



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:

$$\begin{aligned}
Pr_{AA}(v,\kappa,\pi) &= \pi_A * p_{AA}(v) = 0.301 * 0.991 = 0.299 \\
Pr_{TT}(v,\kappa,\pi) &= \pi_T * p_{TT}(v) = 0.269 * 0.989 = 0.266 \\
Pr_{GG}(v,\kappa,\pi) &= \pi_G * p_{GG}(v) = 0.121 * 0.988 = 0.120 \\
Pr_{AA}(v,\kappa,\pi) &= \pi_A * p_{AA}(v) = 0.301 * 0.991 = 0.299 \\
Pr_{CC}(v,\kappa,\pi) &= \pi_C * p_{CC}(v) = 0.308 * 0.990 = 0.305 \\
Pr_{AC}(v,\kappa,\pi) &= \pi_A * p_{AC}(v) = 0.301 * 0.003 = 0.001 \\
Pr_{AC}(v,\kappa,\pi) &= \pi_A * p_{AC}(v) = 0.301 * 0.003 = 0.001 \\
Pr_{AC}(v,\kappa,\pi) &= \pi_A * p_{AC}(v) = 0.301 * 0.003 = 0.001 \\
Pr_{CA}(v,\kappa,\pi) &= \pi_C * p_{CA}(v) = 0.308 * 0.003 = 0.001 \\
Pr_{AA}(v,\kappa,\pi) &= \pi_A * p_{AA}(v) = 0.301 * 0.991 = 0.299
\end{aligned}$$

$$L = \prod_{site=1}^{N} Pr_{site} = 2.485E-16$$

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

$$\ell = \sum_{site=1}^{N} \ln(Pr_{site}) = -35.931$$

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`: data frame 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:

$$\ell = -3104.940$$

We can now fix all parameters except the transition transversion rate ratio κ and look for the maximum log-likelihood (Figure 1) around $\hat{\kappa}$ we estimated.

Figure 1: Log-likelihood curve of the proposed model

The estimate of $\hat{\kappa} = 1.890$ is very close to the maximum but there is a transition transversion rate ratio which gives a slightly better likelihood:

$$\hat{\kappa} = 1.932$$
$$\ell = -3104.923$$

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 your validation.

To simulate two aligned sequences over a distance of $v = 0.01$ and the previously defined substitution rate matrix $\boldsymbol{Q}$ it is assumed that the substitutions happen independently across sites.
10 pairwise alignments of length 1140 nt were simulated.

To validate the model, we can compare the probabilities for each of the 16 site patterns with the observed frequencies from the simulated alignments (Table 3).
The calculated probabilities are very close to the observed frequencies which verifies the model.

Table 3: Pattern frequencies and probabilities

pattern	frequency	probability
AA	0.3004	0.2987
AC	0.0011	0.0010
AG	0.0013	0.0007
AT	0.0014	0.0009
CA	0.0007	0.0010
CC	0.3041	0.3053
CG	0.0004	0.0004
CT	0.0013	0.0017
GA	0.0005	0.0007
GC	0.0003	0.0004
GG	0.1202	0.1200
GT	0.0003	0.0003
TA	0.0009	0.0009
TC	0.0021	0.0017
TG	0.0001	0.0003
TT	0.2648	0.2660

The log-likelihood of each simulated alignment was computed and the distribution is shown in Figure 2.
Much higher log-likelihoods compared to what was calculated earlier hints that the given distance might not be an accurate assumption for the sequence alignment provided (see Discussion).

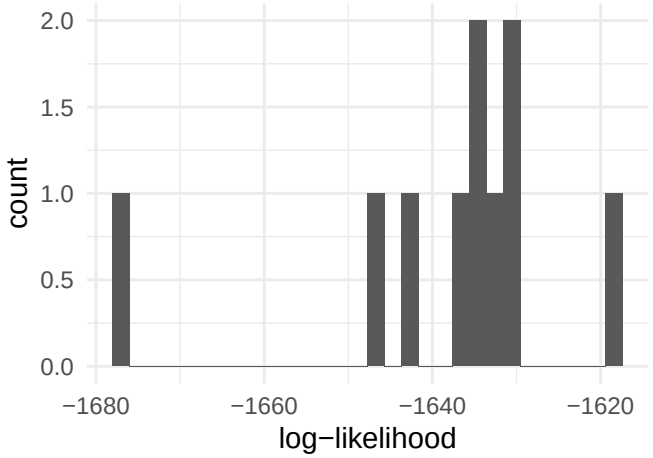


Figure 2: Log-likelihood distribution of simulated sequences

Discussion

It's unclear if the 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 over fitting.

The evolutionary distance given ($v = 0.01$) is much lower compared to the estimated distances $\hat{v}$ (Appendix). Changing the given 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 κ we get following results:

$$\hat{v} = 0.306$$

$$\hat{\kappa} = 2.168$$

with a log-likelihood of $\ell = -2427.557$ which is higher than the previously calculated maximum likelihood.

Appendix

As mentioned, there exist many different established Markov models of nucleotide substitution. Table 4 lists estimates using the given alignment under different models (JC69[7], K80[8], F81[9], HKY85[5]). The results were computed using either formulae where possible[4,10] or maximum likelihood[4].
One should be careful when comparing log-likelihoods under different models with different numbers of free parameters. Log-likelihoods tend to be higher using a model with more parameters.

Table 4: Comparing estimates using different models

Model and method	$\hat{v}$	$\hat{\kappa}$	$(\hat{\pi}_T, \hat{\pi}_C, \hat{\pi}_A, \hat{\pi}_G)$	ℓ
Formulae				
JC69	0.3028			
K80	0.3051	2.1248		
F81	0.3050			
Maximum likelihood				
JC69	0.3028			-2532.34
K80	0.3052	2.1257		-2518.28
F81	0.3023		(0.2677, 0.3200, 0.2998, 0.1126)	-2441.58
HKY85	0.3050	2.1699	(0.2656, 0.3176, 0.3045, 0.1123)	-2426.77

Disclosure

Statistics were done using `R`[11] (3.6.1) and the following packages:

- `tidyverse`[12] (1.3.0)
- `seqinr`[13] (3.6-1)
- `knitr`[14] (1.26)
- `kableExtra`[15] (1.1.0)
- `expm`[16] (0.999-4)

I'd like to thank Teja Turk for helping me apply `stats::optim()` with multiple parameters to be optimized over, showing me LaTeX tricks and for providing moral support.

Notation

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

References

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

7. Jukes, T. H., Cantor, C. R. & others. Evolution of protein molecules. *Mammalian protein metabolism* **3**, 132 (1969).

8. 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).

9. Felsenstein, J. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376 (1981).

10. McGuire, G., Prentice, M. & Wright, F. Improved error bounds for genetic distances from dna sequences. *Biometrics* **55**, 1064–70 (2000).

11. R Core Team. *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing, 2019).

12. Wickham, H. *et al.* Welcome to the tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

13. Charif, D. & Lobry, J. SeqinR 1.0-2: A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. in *Structural approaches to sequence evolution: Molecules, networks, populations* (eds. Bastolla, U., Porto, M., Roman, H. & Vendruscolo, M.) 207–232 (Springer Verlag, 2007).

14. Xie, Y. Knitr: A comprehensive tool for reproducible research in R. in *Implementing reproducible computational research* (eds. Stodden, V., Leisch, F. & Peng, R. D.) (Chapman; Hall/CRC, 2014).

15. Zhu, H. *KableExtra: Construct complex table with 'kable' and pipe syntax.* (2019).

16. Goulet, V. *et al. Expm: Matrix exponential, log, 'etc'.* (2019).