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 initial state distribution (π) 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].

?initial state distribution = steady-state distribution?

Estimations to parameterize the model

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.

?These are the initial state distribution probabilities π_i or equilibrium distribution π?
?what's the difference between π and $\hat{\pi}$

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 equilibrium distribution π.

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

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 numbers of the 16 possible combinations of the sequence alignment are shown in Table 2 (from mouse in rows to human in columns).

Table 2: Nucleotide comparisons (numbers)

	T	C	A	G
T	232	72	20	3
C	32	249	27	4
A	19	59	264	19
G	3	11	15	111

We get the frequencies (Table 3) when the numbers are divided by the total length (1140 nt).

Table 3: 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 3.
The proportion of different sites ($\hat{p} = S + V = 0.2491$) usually underestimates the amount of substitutions which occured.
To get a better estimate, we can apply the distance formulae for the K80 model which consideres different transitions- transversion-rates[4].
The substitution rate matrix (Q) of the K80 model is

$$Q = \begin{bmatrix} . & \alpha & \beta & \beta \\ \alpha & . & \beta & \beta \\ \beta & \beta & . & \alpha \\ \beta & \beta & \alpha & . \end{bmatrix}.$$

The estimate of distance ($\hat{d}$) and transition/transversion rate ratio ($\hat{\kappa}$) can now be calculated as follows:

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

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

(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) we're given that the distance of the two sequences from human and mouse is $\hat{d} = 0.01$. This means that the predicted actual mutations between the two sequences is 0.01 per position.
?how do we use that information?

The likelihood at a position is defined as the sum of the probabilities of all possible ancestral nucleotides. For this pairwise alignment we assume one internal node (one common ancestor) which gives us 4 possibilities.

For this example, likelihood is defined as the product of the likelihoods
of each of the 10 positions:

$$L = \prod_{j=1}^{N} L_{(j)}$$

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

$$\ell = \sum_{j=1}^{N} ln(L_{(j)})$$

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?

Implementing a simulation

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

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.