Mutation models I: basic nucleotide sequence mutation models

Peter Beerli

September 21, 2017

Mutations are irreversible changes in the DNA. This changes may be introduced by chance, by chemical agents, or radioactive decay.

1 Types of mutations

Many mutations involve larger sections of the DNA, these increase or decrease the DNA content

- Duplication: a segment of a chromosome is duplicated, resulting in potential overdoses of proteins that are coded in the duplicated regions, this mechanism is thought to be one of the driving forces of evolution because one of the two copies can now take another function.
- Deletions: a section of a chromosome is deleted, these section can be small (1 base) or large (many 10,000 bases)
- Translocation: one segment of a chromosome is exchanged with another segment of another chromosome.
- Inversion: a chromosome section is flipped.
- Slippage: a long stretch of repeated nucleotide patterns is not exactly copied in the DNA strand replication (details in the chapter *mutation models 3.*)
- Point mutations: these affect always a single DNA site. Typically a nucleotide is damaged and during the next replication of the cell that damaged nucleotide is replaced with another nucleotide. Because all coding regions on the DNA are organized in triplets that are then used to synthetize the protein using transfer-RNA and ribosomes, and because these triplets are

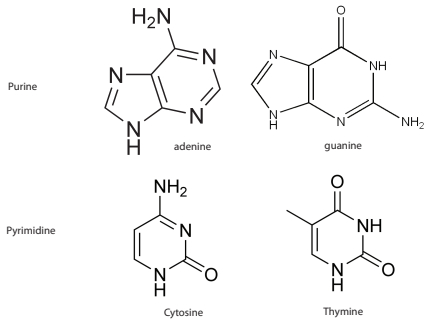


Figure 1: Building blocks of DNA

redundant, there are potentially 4^3 triplet combinations (with four nucleotides A=adenine, C=cytosine, G=guanine, T=thymine [in RNA T is replaced by U=uracile]), but only 20 common aminoacids and 3 STOP codons to mark the end of a coding sequence. We will talk in more detail about codons in *mutation models 2*.

2 Point mutations

2.1 A simple model

DNA is built out 4 nucleotides and a model would need to take into account for mutations that allow to transition from one particular nucleotide to another, in the following section we reduce the problem to very simple model with two states U and Y but this shows the whole complexity of the modelling process. If you wish you could think of this as pUrines (either the nucleotide adenosine [A] or a guanine [G]) or pYrimidines (either cytosine [C] or thymine [T]) (Figure 1). We have the states and substitution rate μ for the substitution rate from U to Y and the transition rate μ from Y to U (Figure 2). This is a very simple model. We could think of introducing a different rate

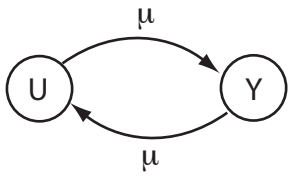


Figure 2: Simple model with two states and one mutation rate

for going from Y to U , but will refrain to do so for this outline. We will see later that there are

models that set back-mutation to zero. The most common of these is the infinite sites mutation model. It will be discussed in chapter *mutation models 3*.

2.1.1 Discrete time

The model shown in Figure 2 assumes that time is discrete and that we evaluate the transition from one state in time i to the same or other state in time $i+1$, where the allele U is at risk to mutate to Y with rate μ or to stay in U with rate $1-\mu$. The same logic applies to Y where Y changes to U with rate μ and stays in Y with rate $1-\mu$. We can express this as a transition matrix

$$R = \begin{pmatrix} 1-\mu & \mu \\ \mu & 1-\mu \end{pmatrix} \tag{1}$$

often expressed like this

$$R_* = \begin{pmatrix} -\mu & \mu \\ \mu & -\mu \end{pmatrix} \tag{2}$$

$$R = I + R_* \tag{3}$$

where I is the Identity matrix. When we start with a specific state, say U, then we can evaluate in what state we will be in the next time-click. Running this process for many steps creates a Markov chain, in which the next state is only dependent on the current state and on the transition matrix that formulates the probabilities of change from U to Y or from Y to U or the probability of no change.

2.1.2 Stationary distributions

Unspoken assumptions of the above framework are:

- The Markov chain is *irreducible*: we can reach every state from any other, in our example we can go from U to Y and from Y to U. If we would set one of the transition rates to 0 then our framework will have a problem.
- The chain is aperiodic, we never go into a loop that cycles forever only in a subset of solutions. As long as μ and ν in our sample are not zero we will visit every state.

- All states of the chain are *ergodic*[1]. When we run the chain infinitely long every state has a non-zero probability π_i. So we can say that the rate matrix R has stationary distribution Π (diagonal matrix)

$$\lim_{n->\infty}(R^n) = \Pi \text{ or } \lim_{n->\infty}(R^n)_{ij} = \pi_i \tag{4}$$

2.1.3 Divergence matrix

The matrix R gives the conditional probabilities:

$$R^k_{ij} = \text{Prob(in state } j \text{ after } k \text{ ticks}|\text{state } i)$$

the matrix $X(k)$ is defined as the divergence matrix

$$X(k)_{ij} = \text{Prob(in state } j \text{ after } k \text{ ticks AND state } i)$$

We assume that the initial state was sampled from the stationary distribution (the process is already at equilibrium). Then

$$X(k)_{ij} = \pi_i(R^k)_{ij} \text{or } X(k) = \Pi R^k$$

Typically we would think of Π as a matrix with the stationary frequencies on the diagonal and zero anywhere else. There is some inconsistency between the graph theory and the phylogenetic literature about rate matrices and sometimes the stationary frequencies are considered to be part of the rate matrix R.

2.2 Time reversible models

We say that the Markov chain is time reversible if the divergence matrix $X(k)$ is symmetric for all k so that $X(t)_{ij} = X(t)_{ji}$. Therefore we can calculate probabilities on trees without bothering whether this is forward or backward in time and any calculations can be made on unrooted trees. This assumptions makes the calculation of probabilities of a pair of nodes A and B that have a root C easy as the probabilities do not depend on the actual time of a node but only on the total branch length between A and B, so it is the simple addition of length A-C and C-B.

[1]ergodic – relating to or denoting systems or processes with the property that, given sufficient time, they include or impinge on all points in a given space and can be represented statistically by a reasonably large selection of points.

2.2.1 Calculation of probabilities for continuous time [see also appendix]

We discussed a discrete model that would force us to know the number and duration of the time 'ticks', both is typically not known. instead of having fixed time events we can assume that the events are drawn from a Poisson distribution, with probability

$$\text{Prob}(k \text{ events}|t,\mu) = e^{-\mu t}\frac{(\mu t)^k}{k!} \tag{5}$$

the expected number of events is μt and we call μ the expected number of events per unit time – biologists would think of this as the *mean instantaneous substitution rate*.

Let $P(t)_{ij}$ be the probability that being in state j at time t when started at time zero in state i.

$$P(t) = \begin{pmatrix} \text{Prob}(U \to U|t) & \text{Prob}(U \to Y|t) \\ \text{Prob}(Y \to U|t) & \text{Prob}(Y \to Y|t) \end{pmatrix} \tag{6}$$

$$P(t) = \sum_{k=0}^{\infty} R^k \ \text{Prob}(k \text{ events } |t,\mu) \tag{7}$$

$$= \sum_{k=0}^{\infty} R^k \ e^{-\mu t}\frac{(\mu t)^k}{k!} \tag{8}$$

$$= e^{\mu t}\sum_{k=0}^{\infty} R^k \ \frac{(\mu t)^k}{k!} \tag{9}$$

The sum in formula 9 has the same form as the series approximation of the matrix exponentiation

$$e^A = \sum_{k=0}^{\infty}\frac{A^k}{k!} \tag{10}$$

and so we get

$$P(t) = e^{-\mu t}e^{R\mu t} \tag{11}$$

$$= e^{-\mu t I}e^{R\mu t} \qquad \text{I is the identity matrix} \tag{12}$$

$$= e^{(R-I)\mu t} \tag{13}$$

$$= e^{Q\mu t} \tag{14}$$

where $Q = R - I$ is the is the instantenous substitution rate matrix.

2.2.2 Matrix exponentiation

Matrix exponentiation is not all that easy, the formula 10 does converge only very slowly and often is not very useful. A better approach for some square and real matrices is using an Eigen decomposition because

$$e^A = P^{-1} e^D P \tag{15}$$

$$e^D = \begin{pmatrix} e^{d_{11}} & 0 & ... & \\ 0 & e^{d_{22}} & 0 & ... \\ ... & & & \\ 0 & ... & 0 & e^{d_{nn}} \end{pmatrix} \tag{16}$$

where P are the Eigenvectors and D is the diagonal matrix of the Eigenvalues. Matrix exponentiation is a very common problem in scientific computing and many different approaches exist, a good overview is given by Moler (2003) in *Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later* (SIAM REVIEW 2003 Society for Industrial and Applied Mathematics Vol. 45, No. 1, pp. 3000).

2.2.3 Substitution matrix for our simple 2-state case

With this simple case we can solve the equation analytically. Using MATHEMATICA we find the probability transition matrix

$$\begin{pmatrix} \text{Prob}(U \to U|t) & \text{Prob}(U \to Y|t) \\ \text{Prob}(Y \to U|t) & \text{Prob}(Y \to Y|t) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left(1 + e^{-2t\mu}\right) & \frac{1}{2} - \frac{1}{2}e^{-2t\mu} \\ \frac{1}{2} - \frac{1}{2}e^{-2t\mu} & \frac{1}{2}\left(1 + e^{-2t\mu}\right) \end{pmatrix} \tag{17}$$

Some prefer to name this matrix *substitution matrix* because in biology the term *transition* is occupied for the transition from a nucleotide A to G, G to A or from C to T, T to C.

2.3 Nucleotide models

Many models are possible and only few have names, we show not even all of these. Most commonly used are Jukes-Cantor, Kimura-2 parameter model, Felsenstein 81, Hasegawa-Kishino-Yano, Tamura-Nei, and General time reversible models. The differences between these models stems from the fact that they allow for different numbers of parameters. We typically order the substitution matrix

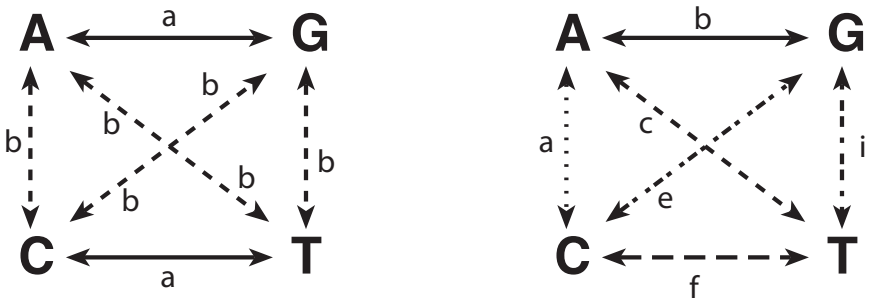


Figure 3: Left: Kimura 2-parameter model. Rates differ between transitions and transversion. Right: General time reversible model. Rates are symmetrical but all pairs of substitution rates are different from each other.

	A	C	G	T
A	-	a	b	c
C	d	-	e	f
G	g	h	-	i
T	k	l	m	-

and allow for maximally 12 modifiers of the overall mutation rate μ, labelled a to m. For an alternative ordering of the mutation matrix see Felsenstein's book, he orders A, G, C, T.

Often in the literature there is no distinction between the divergence matrix X and the instantaneous rate matrix Q. In the following discussion we present the equilibrium frequencies as part of the formula. So that we express

$$X(t) = \Pi Q \mu t \tag{18}$$

2.3.1 General time reversible model (GTR)

$$Q = \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & \mu f\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix}$$

GTR is the most complex time-reversible model with 6 rate parameters $r_{ij} = a, b, c, d, e, f$ and base frequencies π_A, π_C, π_G; π_T is not a real parameter because the base frequencies have to add up to 1. Both, r_{ij} and the base frequencies , form the rates in the Q matrix.

2.3.2 Jukes-Cantor

Jukes-Cantor (JC) allows for a single parameter and has a transition matrix

$$Q = \begin{pmatrix} -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{pmatrix}$$

The base frequencies $\pi_A, \pi_C, \pi_G, \pi_T$ are all the same and 0.25. There are only two types of changes possible, either one does not change or one changes. This results in two probabilities:

$$\text{Prob}(t)_{ii} = \frac{1}{4} + \frac{3}{4}e^{-\mu t} \tag{19}$$

$$\text{Prob}(t)_{ij} = \frac{1}{4} - \frac{1}{4}e^{-\mu t} \tag{20}$$

Compared to the most complex model, GTR, JC is setting all parameters a to f to 1 and all base frequencies to the same value.

2.3.3 Kimura's 2-parameter model

Kimura's two-parameter (K2P) allows for two parameters, different rates for transitions and transversion but still assumes equal base frequencies and has a transition matrix

$$Q = \begin{pmatrix} -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu & \frac{1}{4}\mu\kappa & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu & \frac{1}{4}\mu\kappa \\ \frac{1}{4}\mu\kappa & \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu\kappa & \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) \end{pmatrix}$$

The base frequencies $\pi_A, \pi_C, \pi_G, \pi_T$ are all the same and 0.25. the parameter κ represents the transition bias, when $\kappa = 1$ then the models reduces to the JC model. If the importance of transitions and transversions are rated equally then κ should be two because there are twice as many transversions as transitions. There are three types of changes possible:

$$\text{Prob}(t)_{ii} = \frac{1}{4} + \frac{3}{4}e^{-\mu t} \tag{21}$$

$$\text{Prob}(t)_{ij,\text{Transition}} = \frac{1}{4} + \frac{1}{4}e^{-\mu t} + \frac{1}{2}e^{-\mu t(\frac{\kappa+1}{2})} \tag{22}$$

$$\text{Prob}(t)_{ij,\text{Tranversion}} = \frac{1}{4} - \frac{1}{4}e^{-\mu t} \tag{23}$$

Using GTR we can express the K2P model as $a = c = d = f = 1$ and $b = e = \kappa$.

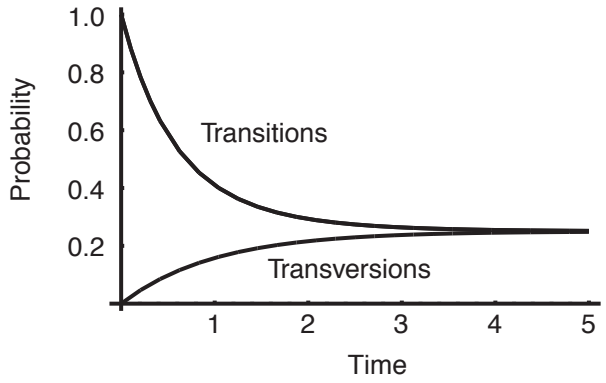


Figure 4: Using different rates for transitions and transversions results in different substitution probabilities.

2.3.4 Hasegawa-Kishino-Yano 1985 and Felsenstein 1984

Hasegawa-Kishino-Yano and Felsenstein relaxed the K2P model and allowed for unequal base frequencies

$$Q_{HKY} = \begin{pmatrix} -\mu(\kappa\pi_G + \pi_Y) & \mu\pi_C & \mu\kappa\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\kappa\pi_T + \pi_R) & \mu\pi_G & \mu\kappa\pi_T \\ \mu\kappa\pi_A & \mu\pi_C & -\mu(\kappa\pi_G + \pi_Y) & \mu\pi_T \\ \mu\pi_A & \mu\kappa\pi_C & \mu\pi_G & -\mu(\kappa\pi_C + \pi_R) \end{pmatrix}$$

where $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$. There are three types of changes possible:

$$\text{Prob}(t)_{jj} = \pi_j + \pi_j \left(\frac{1}{\Pi_j} - 1\right) e^{-\mu t} + \left(\frac{\Pi_j - \pi_j}{\Pi_j} - 1\right) e^{-\mu t\alpha} \tag{24}$$

$$\text{Prob}(t)_{ij,\text{Transition}} = \pi_j + \pi_j \left(\frac{1}{\Pi_j} - 1\right) e^{-\mu t} - \left(\frac{\pi_j}{\Pi_j} - 1\right) e^{-\mu t\alpha} \tag{25}$$

$$\text{Prob}(t)_{ij,\text{Tranversion}} = \pi_j(1 - e^{-\mu t}) \tag{26}$$

where $\Pi_j = \pi_A + \pi_G$ if the base j is a purine, and $\Pi_j = \pi_C + \pi_T$ is a pyrimidine. The parameter α is different between the HKY and the F84 model:

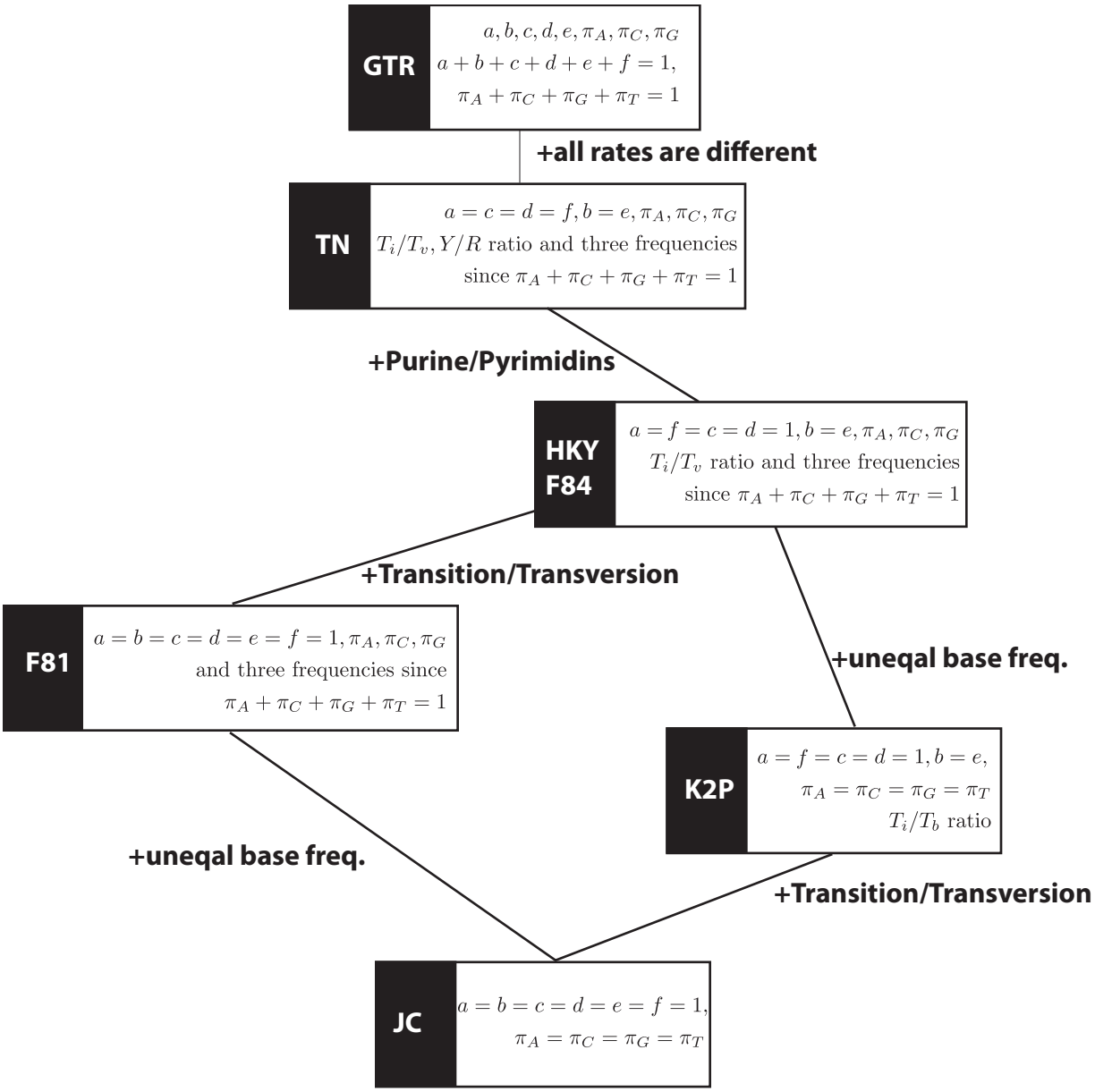
- $\alpha = 1 + \Pi_j(\kappa - 1)$ for the HKY model
- $\alpha = \kappa + 1$ for the F84 model

GTR expresses the HKY model setting $a = c = d = f = 1$ and $b = e = \kappa$ and unequal base frequencies.

2.3.5 Summary of nucleotide models

Using this simplified Q matrix we express all models

$$Q = \begin{pmatrix} -(a\pi_C + b\pi_G + c\pi_T) & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & -(a\pi_A + d\pi_G + e\pi_T) & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & -(b\pi_A + d\pi_C + f\pi_T) & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & -(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix} \tag{27}$$



2.3.6 Other models

According to Huelsenbeck and Ronquist (2005) there are 203 time-reversible models. It is not difficult to generate them and use the matrix-exponentiation method to get probability estimates, whether these models are all useful is a difficult question. Very few attempts to work with models that are not time-reversible were made [JF:210].

2.4 Reducing constraints

2.4.1 Rate variation among sites

We assumed that each site has the same substitution rate. this can be relaxed when we treat the mutation rate as a random variable with appropriate distribution. Currently almost all programs use the gamma distribution. Its density function with parameters $\alpha, \beta > 0$ is

$$f(r|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} r^{\alpha-1} e^{-r/\beta} \tag{28}$$

$$\Gamma(a) = \int_0^{\infty} \text{Prob}(t|r) f(r|\alpha,\beta) dr \tag{29}$$

We think of this mutation rate variation often as two parameters: the mutation rate μ and the rate modifier r. The rate r is the random variable whereas the mutation rate is fixed. for practical purposes we typically assume the mean of r as 1 and so we can simplify the number of parameters of the gamma distribution and reduce β to $1/\alpha$ (the mean of the gamma distribution is $\alpha\beta$. To calculate the rates we use

$$\text{Prob}(t|r) = e^{Q\mu rt} \tag{30}$$

this results in

$$\text{Prob}(t|r) = \int_0^{\infty} \text{Prob}(t|r) f(r|\alpha) dr \tag{31}$$

where we integrate each element of the matrix separately. The integral is time consuming and needs to be done whenever the branch length changes. A faster approximation is to use a discretization of the gamma distribution.

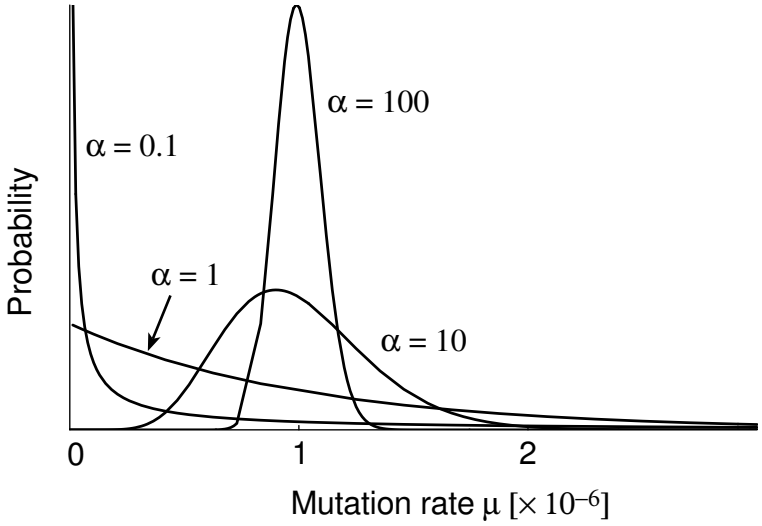


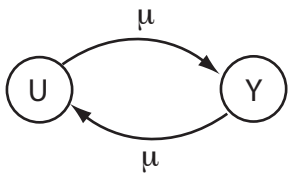
Figure 5: Gamma distributions with several values for α, $\beta = 1/\alpha$

3 Study question

1. What happens when the models are not time reversible? Can you given an example?
2. Change the simple model so that it has a different back mutation, say ν, can you generate the transition matrix Q? It is easiest to think of $\nu = a\mu$.

Appendix: Complete 2-state model example

We have the states and substitution rate μ for the transition rate from U to Y, and the transition rate μ from Y to U.



with transition matrix

$$R = \begin{pmatrix} 1-\mu & \mu \\ \mu & 1-\mu \end{pmatrix} \tag{32}$$

Assuming a Poisson process with the probability that k events are needed to go from state 1 to state 2 in time t is

$$\text{Prob}(k \text{ events}|t,\mu) = e^{-\mu t}\frac{(\mu t)^k}{k!} \tag{33}$$

Writing all transition probabilities into a matrix, we get

$$P(t) = \begin{pmatrix} \text{Prob}(U \to U|t) & \text{Prob}(U \to Y|t) \\ \text{Prob}(Y \to U|t) & \text{Prob}(Y \to Y|t) \end{pmatrix} \tag{34}$$

Inserting and substitution we get

$$P(t) = \sum_{k=0}^{\infty} R^k \text{ Prob}(k \text{ events } |t,\mu) \tag{35}$$

$$= \sum_{k=0}^{\infty} R^k \ e^{-\mu t}\frac{(\mu t)^k}{k!} \tag{36}$$

$$= e^{\mu t}\sum_{k=0}^{\infty} R^k \ \frac{(\mu t)^k}{k!} \tag{37}$$

The sum in formula 37 has the same form as the series approximation of the matrix exponentiation

$$e^A = \sum_{k=0}^{\infty}\frac{A^k}{k!} \tag{38}$$

and so we get

$$P(t) = e^{-\mu t}e^{R\mu t} \tag{39}$$

$$= e^{-\mu t I}e^{R\mu t} \qquad \text{I is the identity matrix} \tag{40}$$

$$= e^{(R-I)\mu t} \tag{41}$$

$$= e^{Q\mu t} \tag{42}$$

where $Q = R - I$ is the is the instantaneous substitution rate matrix. $\exp Q\mu t$ can be solved in many ways, but we use one of the simpler methods. Eigendecomposition

$$A = P\Lambda P^{-1} \tag{43}$$

with the eigenvalues on the diagonal matrix Λ and the eigenvectors P. But before we do that we simplify the exponent, by scaling time by the (expected) mutation rate μ, so that our new scaled time $\tau = \mu t$; therefore we have $\exp(Q\tau)$. Our example decomposes then into

$$Q\tau = (R - I)\tau = \begin{pmatrix} -\mu & \mu \\ \mu & -\mu \end{pmatrix} \tau = \begin{pmatrix} -\mu\tau & \mu\tau \\ \mu\tau & -\mu\tau \end{pmatrix} \tag{44}$$

$$Q\tau = P\Lambda P^{-1} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & -2\mu\tau \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}^{-1} \tag{45}$$

$$= \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & -2\mu\tau \end{pmatrix} \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \tag{46}$$

The exponentiation can be done on the diagonal elements only of Λ, but standard matrix exponentiation $\Lambda^k = \Lambda_1\Lambda_2....\Lambda_k$ results in the same result,

$$P(t) = e^{Q\tau} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} e^0 & 0 \\ 0 & e^{-2\mu\tau} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \tag{47}$$

$$= \begin{pmatrix} \frac{1}{2}\left(1 + e^{-2\tau\mu}\right) & \frac{1}{2} - \frac{1}{2}e^{-2\tau\mu} \\ \frac{1}{2} - \frac{1}{2}e^{-2\tau\mu} & \frac{1}{2}\left(1 + e^{-2\tau\mu}\right) \end{pmatrix} \tag{48}$$