Bioniformatics Makerere
Models for molecular evolution
Timo Koski 2018-08-05

1 Introduction

This lecture applies the results from a previous lecture on Markov chains in continuous time to molecular evolution.

We present some of the many mathematical theories proposed for the study of the mechanisms of DNA evolution. These studies are essential because changes in nucleotide sequences are used both for estimating the rate of evolution and for reconstructing the evolutionary history of organisms.

The idea is that sequences diverge from a common ancestor because mutations occur and some fraction of these mutations are fixed into to the evolving population by selection and chance. A completely general model of this process should take into accoount substitutions, insertion and deletion, duplication and transposition of individual bases and of blocks of bases (Li 1997). The models treated below will only be concerned with substitution, replacement of one base by another.

The perhaps most computationally tractable probabilistic model to consider is the one in which all the instantaneous rates of change are the same for all nucleotides. This is the Jukes-Cantor model. This model and most other models assume that the rate of change is the same at all sites. It is possible to include rate variation, e.g., by assuming that the rates are gamma-distributed random variables (Yang 1993). Felsenstein and Churchill (1996) suggest a hidden Markov model for the variation among sites in rates. The clock hypothesis is more precisely formulated (Zuckerkandl and Pauling 1965, Zuckerkandl et.al. 1971, p. 485) as a simple statistical model, a continuous-time Markov chain, that describes events of substitution occurring at random times with constant rates.

In (Zuckerkandl and Pauling 1965, Zuckerkandl et.al. 1971, loc.cit, Neyman 1970) the disturbances that may result in failure to restore DNA sequences correctly arise according to a *Poisson process*. This means that if $N(t)$, $t \geq 0$, is the number of such disturbances occurring during in a time interval of length t, then $N(t)$ is a discrete random variable having a Poisson distribution with parameter λt. Thus the probability of the occurrencies of

k changes in time t is given by

$$P\left(N(t)=k\right)=e^{-\lambda t}\frac{(\lambda t)^k}{k!}. \tag{1.1}$$

The quantity λ is a non-negative statistical parameter of the model and is called the *intensity*. Then the probability of no disturbance (substitution) is

$$P\left(N(t)=0\right)=e^{-\lambda t}, \tag{1.2}$$

and the probability of at least one disturbance is

$$P\left(N(t)\geq 1\right)=1-e^{-\lambda t}. \tag{1.3}$$

These two last expressions will reappear frequently in the sequel. A property of this model is that failure to copy correctly at a particular nucleotide site is independent of the past history of that site. A Poisson distributed molecular clock can also be derived using population genetics and coalescent theory (Tavaré 1995).

The molecular clock hypothesis for single nucleotide substitutions is next formulated by additional probabilistic assumptions.

1.1 Statistical assumptions on the substitution process

Let us suppose that

1) All nucleotide sites change independently.

2) All nucleotide sites vary equiprobably.

3) The substitution rate is constant over time and in different evolutionary lineages.

4) The base composition is at equilibrium.

5) The conditional probabilities of nucleotide substitutions are the same for all sites and do not change over time.

We are under these premises dealing with a 'space-time' plane, in which a number of independent mutation processes evolve in a parallel way. An attempt to visualize the 'space-time' plane is in Figure 1. In general we

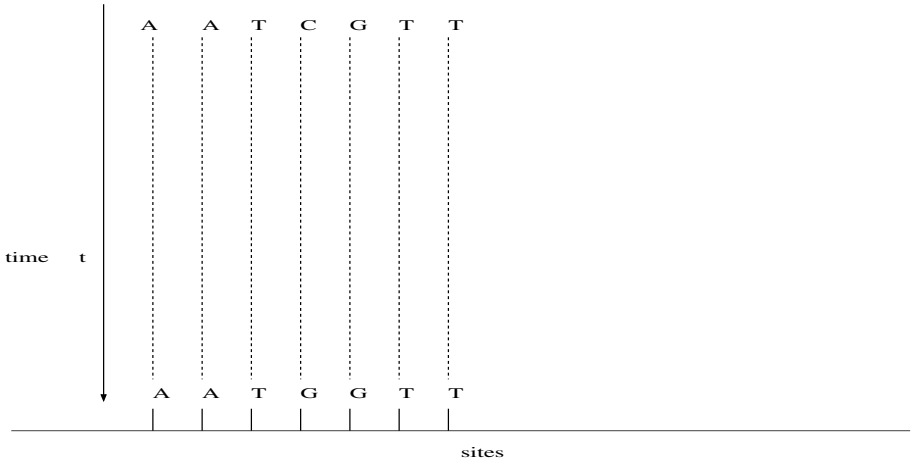


Figure 1: Space-time

should thus introduce a set of independent random processes $X_s = \{X_s(t) \mid t \geq 0\}$, where s denotes the site in a genome. By virtue of the assumptions above it will suffice, at least prior to discussion of phylogenetic inference, to consider only a single process $X = \{X(t) \mid t \geq 0\}$ as a generic model of the substitutions at any site.

2 Summary of some known results

$X = \{X(t) \mid t \geq 0\}$ is a continuous-time Markov chain taking values in $\mathcal{X}$ $= \{A, T, C, G\}$. We denote the generic elements of $\mathcal{X}$ by $j, i, \ldots,$.

The time-homogeneous transition probability is

$$P_{ij}(t) = P\left(X(t) = j | X(0) = i\right)$$

or

$$P_{ij}(t-s) = P\left(X(t) = j | X(s) = i\right)$$

for $j \in \mathcal{X}$. In a matrix form this amounts to

$$\mathbf{P}(t) = \{P_{ij}(t)\}_{i \in \mathcal{X}, j \in \mathcal{X}}.$$

We have a rate matrix $\mathbf{Q}$ given by

$$\lim_{h \downarrow 0} \frac{\mathbf{P}(h) - \mathbf{I}}{h} = \mathbf{Q},$$

and the differential equations

$$\mathbf{P}^{'}(t) = \mathbf{P}(t)\mathbf{Q} \tag{2.1}$$

with the initial conditions

$$P(0) = \mathbf{I}. \tag{2.2}$$

The solution to (2.1) and (2.2) is

$$\mathbf{P}(t) = e^{\mathbf{Q}t} = \sum_{l=0}^{\infty} \frac{t^l}{l!} \mathbf{Q}^l. \tag{2.3}$$

The absolute probability distribution on $\mathcal{X}$ evolves as

$$\mathbf{p}(t) = \mathbf{p}(0)\mathbf{P}(t). \tag{2.4}$$

The stationary distribution π on $\mathcal{X}$ is defined as

$$\pi = \pi \mathbf{P}(t)$$

and found from

$$\pi \mathbf{Q} = \mathbf{0}.$$

Here $\mathbf{0}$ is matrix of zeros.

3 Reversibility

We need the additional assumption of *reversibility*, which means that

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t) \quad \text{for all } t,\ i, j\ \mathcal{X}. \tag{3.5}$$

Proposition 3.1 *A continuous time ergodic and stationary Markov chain is reversible if and only if*

$$\pi_i q_{ij} = \pi_j q_{ji} \quad \textit{for all } i, j. \tag{3.6}$$

Proof: The necessity of (3.6) follows from the assumption of reversibility (3.5) by taking appropriate limits. To prove sufficiency, we introduce the diagonal matrix $\mathbf{\Pi}$, which has an equilibrium distribution on the main diagonal, or

$$\mathbf{\Pi} = \begin{pmatrix} \pi_1 & 0 & 0 & \dots & 0 \\ 0 & \pi_2 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & \pi_{|\mathcal{X}|} \end{pmatrix}.$$

This is an invertible matrix, as all $\pi_i > 0$ for ergodic Markov chains in continuous time. Then we can write the conditions (3.6) in matrix form as

$$\mathbf{\Pi Q} = \mathbf{Q}^T \mathbf{\Pi} \tag{3.7}$$

^{T} is the transpose of a matrix. This is equivalent to

$$\Leftrightarrow \mathbf{\Pi Q \Pi}^{-1} = \mathbf{Q}^T. \tag{3.8}$$

Next we compute

$$\mathbf{\Pi P}(t)\mathbf{\Pi}^{-1} = \mathbf{\Pi} e^{t\mathbf{Q}} \mathbf{\Pi}^{-1} = e^{t\mathbf{\Pi Q \Pi}^{-1}}$$

by a known property of the matrix exponential (Braun 1993, p. 154, Brockett 1970 p. 32). The equality in the right hand side of the equivalence in (3.8) gives

$$= e^{t\mathbf{\Pi Q \Pi}^{-1}} = e^{t\mathbf{Q}^T} = \left(e^{t\mathbf{Q}}\right)^T = \mathbf{P}(t)^T,$$

where another elementary property of the matrix exponential has been used. In other words we have obtained that

$$\mathbf{\Pi P}(t)\mathbf{\Pi}^{-1} = \mathbf{P}(t)^T \Leftrightarrow \mathbf{\Pi P}(t) = \mathbf{P}(t)^T \mathbf{\Pi}, \tag{3.9}$$

and the equality in the right hand side of this equivalence is clearly nothing else but the matrix form of the reversibility conditions (3.5). ∎

3.1 The Most General Time-Reversible Model (GTR)

Tavaré (1986) and several others have suggested the most general reversible model (GTR) as

$$\mathbf{Q} = \begin{pmatrix} * & \mu a\pi_2 & \mu b\pi_3 & \mu c\pi_4 \\ \mu a\pi_1 & * & \mu d\pi_3 & \mu e\pi_4 \\ \mu b\pi_1 & \mu d\pi_2 & * & \mu f\pi_4 \\ \mu c\pi_1 & \mu e\pi_2 & \mu f\pi_3 & * \end{pmatrix}, \tag{3.10}$$

where the diagonal intensities $*$ are obtained by the standard rule. The rows and columns correspond to A,C,T,G in this order. Here μ is the average instantaneous substitution rate, which is modulated by the parameters a, b, c, d, e, f, which correspond to the substitution rates from one base to another. The π_i are collected in

$$\pi = (\pi_1, \pi_2, \pi_3, \pi_4),$$

which must be an equilibrium distribution, and is arbitrary in the sense that it can be preassigned. It is immediate from proposition that $\mathbf{Q}$ in (3.10) indeed defines a reversible model.

In other words, it is *assumed that the rate of change to each base is proportional to the equilibrium probability of that base.* More specialized reversible models are obtained imposing restrictions on $(\pi_1, \pi_2, \pi_3, \pi_4)$ and/or a, b, c, d, e, f.

3.2 The Felsenstein Model (F81) of Evolutionary Change in Nucleotide Sequences

If we choose $a = b = c = d = e = f = 1$ in the GTR generator $\mathbf{Q}$ in (3.10) and set

$$u_i = \mu\pi_i, \quad i = 1, \ldots, 4$$

then the generator is of the form introduced in (3.13), and also known under the acronym **F81** for Felsenstein (1981), i.e.,

$$\mathbf{Q} = \begin{pmatrix} -(\mu - u_1) & u_2 & u_3 & u_4 \\ u_1 & -(\mu - u_2) & u_3 & u_4 \\ u_1 & u_2 & -(\mu - u_3) & u_4 \\ u_1 & u_2 & u_3 & -(\mu - u_4) \end{pmatrix}. \tag{3.11}$$

Then we can write

$$\pi = \left(\frac{u_1}{\mu}, \frac{u_2}{\mu}, \frac{u_3}{\mu}, \frac{u_4}{\mu} \right), \tag{3.12}$$

and also check $\pi \mathbf{Q} = \mathbf{0}$.

We have written earlier $\mu = u$, or

$$\mathbf{Q} = \begin{pmatrix} -(u-u_1) & u_2 & u_3 & u_4 \\ u_1 & -(u-u_2) & u_3 & u_4 \\ u_1 & u_2 & -(u-u_3) & u_4 \\ u_1 & u_2 & u_3 & -(u-u_4) \end{pmatrix}, \tag{3.13}$$

where

$$u = u_1 + u_2 + u_3 + u_4. \tag{3.14}$$

If X is a continuous time Markov chain with the generator $\mathbf{Q}$ in (3.13), then

$$\pi \mathbf{Q} = \mathbf{0}$$

has the solution (the stationary distribution)

$$\pi = \left(\frac{u_1}{u}, \frac{u_2}{u}, \frac{u_3}{u}, \frac{u_4}{u} \right). \tag{3.15}$$

The rate of change R is

$$R = \lim_{h \downarrow 0} \frac{P\left(X(t+h) \neq X(t)\right)}{h} = u \left(1 - \sum_{i \in \mathcal{X}} \pi_i^2 \right). \tag{3.16}$$

where u is given in (3.14). We set

$$\mathbf{A} = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 & \pi_4 \\ \pi_1 & \pi_2 & \pi_3 & \pi_4 \\ \pi_1 & \pi_2 & \pi_3 & \pi_4 \\ \pi_1 & \pi_2 & \pi_3 & \pi_4 \end{pmatrix}.$$

we have

$$\mathbf{P}(t) = e^{\mathbf{Q}t} = e^{-ut}\mathbf{I} + \mathbf{A}\left(1 - e^{-ut}\right), \quad \text{for } \mathbf{Q} \text{ in (3.13)}. \tag{3.17}$$

If we write this elementwise, we get

$$P_{ij}(t) = e^{-ut}\delta_{i,j} + \left(1 - e^{-ut}\right)\pi_j, \tag{3.18}$$

where $\delta_{i,j}$ is the Kronecker delta defined by

$$\delta_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases} \tag{3.19}$$

3.3 Separation of Species

We are assuming $\mathbf{Q}$ in (3.13). Let us now suppose that we have two continuous-time Markov chains X and Y, with the same generator $\mathbf{Q}$ in (3.13), assuming reversibility, and such that

$$X(0) = Y(0),$$

but evolving independently thereafter. Then

$$P\left(X(t) = i, Y(t) = j\right) = \pi_i P_{ij}(2t) =$$

$$= \begin{cases} \pi_i \left(1 - e^{-2ut}\right) \pi_j & i \neq j, \\ \pi_i e^{-2ut} + \pi_i \left(1 - e^{-2ut}\right) \pi_j & i = j. \end{cases} \tag{3.20}$$

Then

$$P\left(X(t) \neq Y(t)\right) = \left(1 - e^{-2ut}\right) \left(1 - \sum_{i \in \mathcal{X}} \pi_i^2\right). \tag{3.21}$$

Note that since the two sequences are independent, the probability of them having different bases is

$$\pi_i \pi_j, i \neq j.$$

Also $\sum_{i \neq j} \pi_i \pi_j = (1 - \sum_{i \in \mathcal{X}} \pi_i^2)$ is thus clearly the stationary probability of difference in any sequence position between X and Y.

4 Jukes-Cantor Model of Evolutionary Change in Nucleotide Sequences

The Jukes-Cantor generator is a special case of (3.13) and is of the form

$$\mathbf{Q} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}. \tag{4.1}$$

The dimension of α could be

$$[\alpha] = \text{ substitutions/site/year }.$$

and is perhaps of order 10^{-4}. We get from (3.15)

$$\pi = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right). \tag{4.2}$$

Hence we see that the Jukes-Cantor model is reversible. This yields from (3.21) that

$$\begin{aligned} P\left(X(t) \neq Y(t)\right) &= \left(1 - \frac{1}{4}\right)\left(1 - e^{-8\alpha t}\right) \\ &= \frac{3}{4}\left(1 - e^{-8\alpha t}\right). \end{aligned} \tag{4.3}$$

This is the expression (13.31) in Ewens and Grant.

If we now set $p = P\left(X(t) \neq Y(t)\right)$, we obtain from (4.3)

$$\alpha t = \frac{1}{8}\log\left(1 - \frac{4}{3}p\right). \tag{4.4}$$

Hence if p is estimated by evolutionary sequence data as $\widehat{p}$, then

$$\widehat{\alpha t} = \frac{1}{8}\log\left(1 - \frac{3}{4}\widehat{p}\right).$$

Next we find the standard form of the Jukes-Cantor formula. Let us introduce, using the rate of change in (3.16) and $u = 4\alpha$ from (3.14)

$$d \stackrel{\text{def}}{=} 2tR = 2t \cdot u \cdot \left(1 - \sum_{i \in \mathcal{X}} \pi_i^2\right) = 2t(4\alpha)\frac{3}{4}$$

$$\Leftrightarrow$$

$$d = (3\alpha)2t.$$

- Here we note that d is the expectation of a $\mathrm{Po}(3\alpha(2t))$ -distributed random variable. The fact that the intensity $q_i = -3\alpha$ is independent of i in the Jukes-Cantor generator (4.1), implies that if an event is defined as the substitution of one nucleotide for another, the number of times the event occurs up to time t follows a Poisson distribution $\mathrm{Po}(3\alpha t)$. Hence, the expected distance between two independent contemporary populations that descended from a common population and are described by the Jukes-Cantor model is like a $\mathrm{Po}(3\alpha(2t))$ -distributed random variable.

- One can explain the appearance of $2t$ by means of 'line-of-descent-and -ascent' argument. The Jukes-Cantor model is reversible, so that the properties of the stochastic process describing any line of descent are the same as that of describing the process in reverse, that is, by considering the corresponding line of ascent. The elapsed time up the line of ascent from one of the two contemporary populations up to the founder population, and then down to the line of descent from the founder population to the other contemporary population, is $2t$, see Figure 2. The Figure can also be read in terms of the Chapman-Kolmogorov semigroup property.

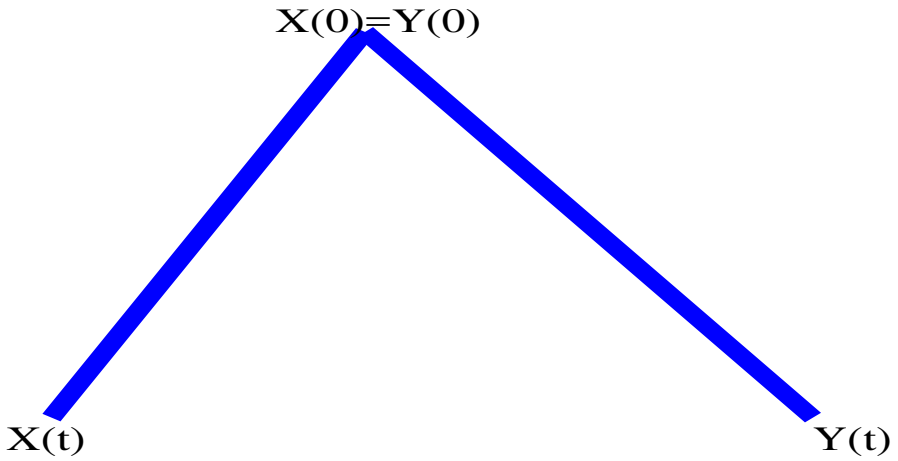


Figure 2: Ascent and descent

Then

$$\alpha t = \frac{d}{6}$$

and this gives (4.3)

$$p = \frac{3}{4}\left(1 - e^{-\frac{4}{3}d}\right). \tag{4.5}$$

or

$$d = -\frac{3}{4}\log\left(1 - \frac{4}{3}p\right).$$

This is the most well known version of the **Jukes-Cantor formula** which thus expresses an evolutionary distance, d, which is the expected number

of substitutions between X and Y as a function of p, which is estimated from data using the observed fractions of positions, where $X(t) \neq Y(t)$. Two examples of uses, of some current interest for the lecturer, are found in (Mougel et. al. (2002), and (Van de Peer et.al. (1993)). In both cases some corrections are introduced, though.

5 The Kimura Two Parameter Model of Evolutionary Change in Nucleotide Sequences

The key assumption of the Jukes-Cantor model that all nucleotide substitutions occur randomly can be quite unrealistic and is mathematically speaking restrictive (but very tractable, as seen above). Nucleotide substitutions can be categorized as transitions and transversions.

- Transitions are substitutions between A and G (purines) or between C and T (pyrimidines).
- Transversions are substitutions between a purine and a pyrimidine (e.g., A changing to C).

Transitions are regarded as more frequent than transversions. Kimura (1980) incorporated this in a two parameter model, in which the rate of transitional substitution at each nucleotide site is per unit time, and the rate of transversions is per unit time.

5.1 The model and its matrix exponential

The Kimura generator of a two parameter model is

$$\mathbf{Q} = \begin{pmatrix} -(\alpha+2\beta) & \beta & \beta & \alpha \\ \beta & -(\alpha+2\beta) & \alpha & \beta \\ \beta & \alpha & -(\alpha+2\beta) & \beta \\ \alpha & \beta & \beta & -(\alpha+2\beta) \end{pmatrix}. \tag{5.1}$$

i.e., as

	A	C	T	G
A	$-(\alpha+2\beta)$	β	β	α
C	β	$-(\alpha+2\beta)$	α	β
T	β	α	$-(\alpha+2\beta)$	β
G	α	β	β	$-(\alpha+2\beta)$

.

The rates for a transition ($= \alpha$) are in this model different from rates for a transversion ($= \beta$). The Jukes-Cantor model is clearly a special case of the Kimura two parameter model, but the Kimura two parameter model is not in general a special case of the Felsenstein model.

- The fact that the intensity $q_i = -(\alpha + 2\beta) < 0$ is independent of i, implies again that for the event of substitution of one nucleotide for another, the number of items the event occurs up to time t follows a Poisson distribution $\mathrm{Po}((\alpha + 2\beta)t)$.

We can write

$$\mathbf{Q} = -(\alpha + 2\beta)\mathbf{I} + \beta\mathbf{E}_1 + \alpha\mathbf{E}_2,$$

where

$$\mathbf{E}_1 = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix},$$

and

$$\mathbf{E}_2 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

One finds easily that

$$\mathbf{E}_1^k = \begin{cases} 2^{k-1}\mathbf{E}_c & k \geq 1, k \text{ even} \\ 2^{k-1}\mathbf{E}_1 & k \geq 1, k \text{ odd}, \end{cases}$$

where

$$\mathbf{E}_c = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix},$$

and

$$\mathbf{E}_2^k = \begin{cases} \mathbf{I} & k \geq 1, k \text{ even} \\ \mathbf{E}_2 & k \geq 1, k \text{ odd}. \end{cases}$$

Then

$$e^{t\mathbf{Q}} = e^{-t(\alpha+2\beta)\mathbf{I}} e^{t\beta\mathbf{E}_1} e^{t\alpha\mathbf{E}_2}$$

and the matrix exponential can be tediously[1] calculated as

$$e^{t\mathbf{Q}} = \begin{pmatrix} f_1(t) & f_2(t) & f_2(t) & f_3(t) \\ f_2(t) & f_1(t) & f_3(t) & f_2(t) \\ f_2(t) & f_3(t) & f_1(t) & f_2(t) \\ f_3(t) & f_2(t) & f_2(t) & f_1(t) \end{pmatrix}, \tag{5.2}$$

where

$$f_1(t) = \frac{1}{4} + \frac{1}{2}e^{-2\alpha t - 2\beta t} + \frac{1}{4}e^{-4\beta t},$$

$$f_2(t) = \frac{1}{4} - \frac{1}{4}e^{-4\beta t},$$

and

$$f_3(t) = \frac{1}{4} - \frac{1}{2}e^{-2\alpha t - 2\beta t} + \frac{1}{4}e^{-4\beta t}.$$

Here $f_3(t)$ is the probability of the event that the initial nucleotide and the nucleotide at t differ by transition and $f_2(t)$ is the probability of the event that the initial nucleotide and the nucleotide at t differ by transversion.

5.2 Stationary distribution, fraction of divergence

Next, if we let $t \to \infty$, we get

$$f_i(t) \to \frac{1}{4}, \quad i = 1, 2, 3,$$

which also shows that the stationary distribution on $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}, \}$ is

$$\pi_{\text{Kimura}} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right). \tag{5.3}$$

Hence we see from (5.2) that the Kimura two parameter model is also reversible in the sense of (3.5). Let us again take two continuous-time Markov chains X and Y, with the same generator $\mathbf{Q}$ in (5.1), and such that

$$X(0) = Y(0) \in \pi,$$

[1] Actually, it is faster to use some software platform of symbolic computation, which Kimura, of course, did not have at his disposal.

but evolving independently thereafter. Then it holds, as has been shown, that

$$P\left(X(t)=i,Y(t)=j\right)=\pi_i P_{ij}(2t)=$$

and that this gives

$$=\begin{cases}\frac{1}{4}\left(\frac{1}{4}-\frac{1}{4}e^{-8\beta t}\right) & i\leftrightarrow j \quad \text{transversion},\\ \frac{1}{4}\left(\frac{1}{4}-\frac{1}{2}e^{-4\alpha t-4\beta t}+\frac{1}{4}e^{-8\beta t}\right) & i\leftrightarrow j \quad \text{transition},\\ \frac{1}{4}\left(\frac{1}{4}+\frac{1}{2}e^{-4\alpha t-4\beta t}+\frac{1}{4}e^{-8\beta t}\right) & i=j.\end{cases} \tag{5.4}$$

Hence we may again explain intuitively $2t$ by means of 'line-of-descent-and-ascent' argument. For the Kimura model it is natural to compute fractions of divergnece both in the case of transversion and transition. By the same computation as above we have

$$p_1=P\left(X(t)\neq Y(t)\mid \text{transition}\right)=\sum_{i\neq j,\text{transition}} P\left(X(t)=i,Y(t)=j\right)$$

$$=\sum_{i\neq j,\text{transition}}\pi_i P_{ij}(2t)=4\frac{1}{4}f_3(2t)$$

$$=\frac{1}{4}-\frac{1}{2}e^{-4\alpha t-4\beta t}+\frac{1}{4}e^{-8\beta t}.$$

And

$$p_2=P\left(X(t)\neq Y(t)\mid \text{transversion}\right)=\sum_{i\neq j,\text{transversion}} P\left(X(t)=i,Y(t)=j\right)$$

$$=\sum_{i\neq j,\text{transversion}}\pi_i P_{ij}(2t)=8\frac{1}{4}f_2(2t)$$

$$=\left(\frac{1}{2}-\frac{1}{2}e^{-8\beta t}\right).$$

Let us set

$$\kappa=4\alpha t+4\beta t,\phi=8\beta t.$$

This gives

$$\kappa=-\log(1-2p_1-p_2),$$

and

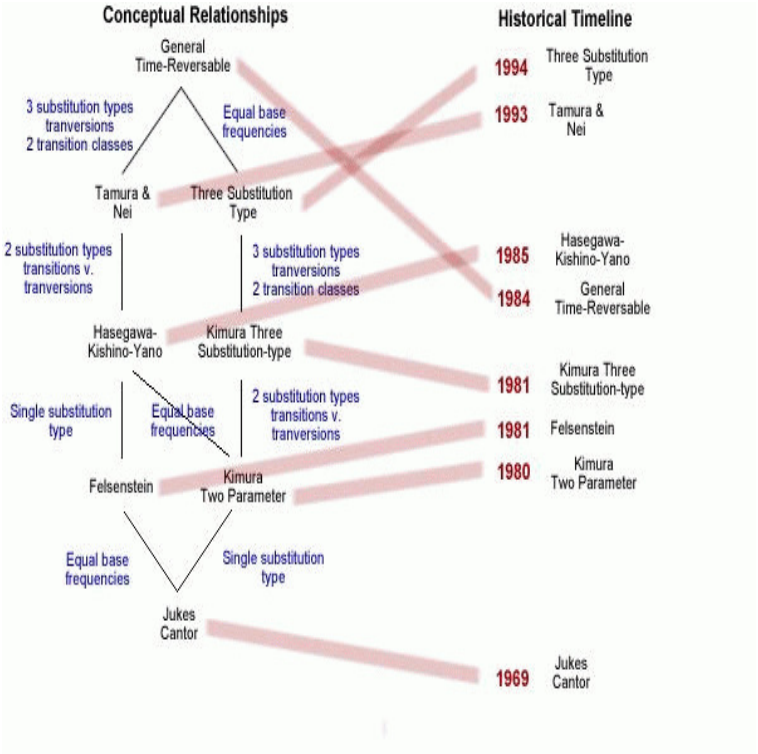
$$\phi=-\log(1-2p_2)$$

as two parameter extensions of the Jukes-Cantor formula.

6 Models of Evolutionary Change in Nucleotide Sequences

In this way many extended models can be obtained. We have a diagram (copied from somewhere in the world wide web) in the appendix giving an outline of the development of this branch of mathematical biology. The book (Li (1997)) mentions many additional variants of generators suggested for molecular evolution.

7 Appendix: A Diagram on Markov Generator Based Theories of Molecular Evolution



8 References

J. Felsenstein (1981): Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *Journal of Molecular Evolution*, 17, pp. 368−376.
J. Felsenstein and G. Churchill (1996): A Hidden Markov Model Approach to Variation Among Sites in Rate of Evolution *Molecular Biology and Evolution*, 17, pp. 93−104.
M. Kimura (1980): A simple method of estimating evolutionary rate in a finite population due to mutational production of neutral and nearly neutral base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Biology,* 16, pp. 111−120.
M. Kimura (1981): Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences, USA,* p. 454−458.
W-H. Li (1997): *Molecular Evolution.* Sinauer Associates, Inc., Publisher.
C. Mougel, J. Thioulouse, G. Perrier and X. Nesme (2002): A mathematical method for determining genome divergence and species delineation using AFLP. *International Journal of Systematic and Evolutionary Microbiology,* 52, pp. 573−586.
S. Tavaré (1986): Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Some Mathematical Questions in Biology. Lectures on Mathematics in the Life Sciences*, 17, The American Mathematical Society, Rhode Island, pp. 57−86.
M. Waterman (1995): *Introduction to Computational Biology. Maps, Sequences and Genomes.* Chapman and Hall, London, et.al.,
Y. Van de Peer, J-M. Neefs, P. de Rijk and R. De Wachter (1993): Reconstructing Evolution from Eukaryotic Small-Ribosomal-Subunit RNA Sequences: Calibration of the Molecular Clock. *Journal of Molecular Evolution*, 37, pp. 221−232.
Z. Yang (1993): Maximum Likelihood Estimation of Phylogeny from DNA Sequences when Substitution Rates Differ over Sites. *Journal of Molecular Evolution*, 10, pp. 1396−1401.
E. Zuckerkandl and L. Pauling (1965): Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*, (V. Bryson and H.J.Vogel (ed.s), Academic Press, New York, pp. 357−366.
E. Zuckerkandl, J. Derancourt and H. Vogel (1971): Mutational Trends and Random Processes in the Evolution of Informational Macromolecules. *Journal of Molecular Biology*, 26, pp. 34−46.

E. Zuckerkandl (1987): On the Molecular Evolutionary Clock. *Journal of Molecular Evolution*, 26, pp. 34−46.