WIKIPEDIA

Models of DNA evolution

A number of different Markov **models of DNA sequence evolution** have been proposed. These substitution models differ in terms of the parameters used to describe the rates at which one nucleotide replaces another during evolution. These models are frequently used in molecular phylogenetic analyses. In particular, they are used during the calculation of likelihood of a tree (in Bayesian and maximum likelihood approaches to tree estimation) and they are used to estimate the evolutionary distance between sequences from the observed differences between the sequences.

Contents

Introduction

These models are phenomenological descriptions of the evolution of DNA as a string of four discrete states. These Markov models do not explicitly depict the mechanism of mutation nor the action of natural selection. Rather they describe the relative rates of different changes. For example, mutational biases and purifying selection favoring conservative changes are probably both responsible for the relatively high rate of transitions compared to transversions in evolving sequences. However, the Kimura (K80) model described below only attempts to capture the effect of both forces in a parameter that reflects the relative rate of transitions to transversions.

Evolutionary analyses of sequences are conducted on a wide variety of time scales. Thus, it is convenient to express these models in terms of the instantaneous rates of change between different states (the Q matrices below). If we are given a starting (ancestral) state at one position, the model's Q matrix and a branch length expressing the expected number of changes to have occurred since the ancestor, then we can derive the probability of the descendant sequence having each of the four states. The mathematical details of this transformation from rate-matrix to probability matrix are described in the mathematics of substitution models section of the substitution model page. By expressing models in terms of the instantaneous rates of change we can avoid estimating a large numbers of parameters for each branch on a phylogenetic tree (or each comparison if the analysis involves many pairwise sequence comparisons).

The models described on this page describe the evolution of a single site within a set of sequences. They are often used for analyzing the evolution of an entire locus by making the simplifying assumption that different sites evolve independently and are identically distributed. This assumption may be justifiable if the sites can be assumed to be evolving neutrally. If the primary effect of natural selection on the evolution of the sequences is to constrain some sites, then models of among-site rate-heterogeneity can be used. This approach allows one to estimate only one matrix of relative rates of substitution, and another set of parameters describing the variance in the total rate of substitution across sites.

DNA evolution as a continuous-time Markov chain

Continuous-time Markov chains

Continuous-time Markov chains have the usual transition matrices which are, in addition, parameterized by time, t. Specifically, if E_1, E_2, E_3, E_4 are the states, then the transition matrix

> $P(t) = \big(P_{ij}(t)\big)$ where each individual entry, $P_{ij}(t)$ refers to the probability that state E_i will change to state E_j in time t.

Example: We would like to model the substitution process in DNA sequences (*i.e.* Jukes–Cantor, Kimura, *etc.*) in a continuous-time fashion. The corresponding transition matrices will look like:

$$P(t) = \begin{pmatrix} p_{\mathrm{AA}}(t) & p_{\mathrm{AG}}(t) & p_{\mathrm{AC}}(t) & p_{\mathrm{AT}}(t) \\ p_{\mathrm{GA}}(t) & p_{\mathrm{GG}}(t) & p_{\mathrm{GC}}(t) & p_{\mathrm{GT}}(t) \\ p_{\mathrm{CA}}(t) & p_{\mathrm{CG}}(t) & p_{\mathrm{CC}}(t) & p_{\mathrm{CT}}(t) \\ p_{\mathrm{TA}}(t) & p_{\mathrm{TG}}(t) & p_{\mathrm{TC}}(t) & p_{\mathrm{TT}}(t) \end{pmatrix}$$

where the top-left and bottom-right 2 × 2 blocks correspond to *transition probabilities* and the top-right and bottom-left 2 × 2 blocks corresponds to *transversion probabilities*.

Assumption: If at some time t_0, the Markov chain is in state E_i, then the probability that at time $t_0 + t$, it will be in state E_j depends only upon i, j and t. This then allows us to write that probability as $p_{ij}(t)$.

Theorem: Continuous-time transition matrices satisfy:

$$P(t + \tau) = P(t)P(\tau)$$

Note: There is here a possible confusion between two meanings of the word *transition*. (i) In the context of *Markov chains*, transition is the general term for the change between two states. (ii) In the context of *nucleotide changes in DNA sequences*, transition is a specific term for the exchange between either the two purines (A ⇔ G) or the two pyrimidines (C ⇔ T) (for additional details, see the article about transitions in genetics). By contrast, an exchange between one purine and one pyrimidine is called a transversion.

Deriving the dynamics of substitution

Consider a DNA sequence of fixed length *m* evolving in time by base replacement. Assume that the processes followed by the *m* sites are Markovian independent, identically distributed and that the process is constant over time. For a particular site, let

$$\mathcal{E} = \{A,\, G,\, C,\, T\}$$

be the set of possible states for the site, and

$$\mathbf{p}(t) = (p_A(t),\, p_G(t),\, p_C(t),\, p_T(t))$$

their respective probabilities at time t. For two distinct $x, y \in \mathcal{E}$, let μ_{xy} be the transition rate from state x to state y. Similarly, for any x, let the total rate of change from x be

$$\mu_x = \sum_{y \neq x} \mu_{xy}\,.$$

The changes in the probability distribution $p_A(t)$ for small increments of time Δt are given by

$$p_A(t+\Delta t) = p_A(t) - p_A(t)\mu_A \Delta t + \sum_{x \neq A} p_x(t)\mu_{xA}\Delta t\,.$$

In other words, (in frequentist language), the frequency of A's at time $+\Delta t$ is equal to the frequency at time t minus the frequency of the *lost* A's plus the frequency of the *newly created* A's.

Similarly for the probabilities $p_G(t)$, $p_C(t)$ and $p_T(t)$. These equations can can be written compactly as

$$\mathbf{p}(t+\Delta t) = \mathbf{p}(t) + \mathbf{p}(t)Q\Delta t\,,$$

where

$$Q = \begin{pmatrix} -\mu_A & \mu_{AG} & \mu_{AC} & \mu_{AT} \\ \mu_{GA} & -\mu_G & \mu_{GC} & \mu_{GT} \\ \mu_{CA} & \mu_{CG} & -\mu_C & \mu_{CT} \\ \mu_{TA} & \mu_{TG} & \mu_{TC} & -\mu_T \end{pmatrix}$$

is known as the *rate matrix*. Note that, by definition, the sum of the entries in each row of Q is equal to zero. It follows that

$$\mathbf{p}'(t) = \mathbf{p}(t)Q\,.$$

For a stationary process, where Q does not depend on time *t*, this differential equation can be solved. First,

$$P(t) = \exp(tQ),$$

where $\exp(tQ)$ denotes the exponential of the matrix tQ. As a result,

$$\mathbf{p}(t) = \mathbf{p}(0)P(t) = \mathbf{p}(0)\exp(tQ)\,.$$

Ergodicity

If the Markov chain is irreducible, *i.e.* if it is always possible to go from a state x to a state y (possibly in several steps), then it is also ergodic. As a result, it has a unique *stationary distribution* $\boldsymbol{\pi} = \{\pi_x,\, x \in \mathcal{E}\}$, where π_x corresponds to the proportion of time spent in state x after the Markov chain has run for an infinite amount of time. In DNA evolution, under the assumption of a common process for each site, the stationary frequencies $\pi_A,\, \pi_G,\, \pi_C,\, \pi_T$ correspond to equilibrium base compositions. Indeed, note that since the stationary distribution $\boldsymbol{\pi}$ satisfies $\boldsymbol{\pi}Q = \mathbf{0}$, we see that when the current distribution $\mathbf{p}(t)$ is the stationary distribution $\boldsymbol{\pi}$ we have

$$\mathbf{p}'(t) = \mathbf{p}(t)Q = \boldsymbol{\pi}Q = 0\,.$$

In other words, the frequencies of $p_A(t),\, p_G(t),\, p_C(t),\, p_T(t)$ do not change.

Time reversibility

Definition: A stationary Markov process is *time reversible* if (in the steady state) the amount of change from state x to y is equal to the amount of change from y to x, (although the two states may occur with different frequencies). This means that:

$$\pi_x \mu_{xy} = \pi_y \mu_{yx}$$

Not all stationary processes are reversible, however, most commonly used DNA evolution models assume time reversibility, which is considered to be a reasonable assumption.

Under the time reversibility assumption, let $s_{xy} = \mu_{xy}/\pi_y$, then it is easy to see that:

$$s_{xy} = s_{yx}$$

Definition The symmetric term s_{xy} is called the *exchangeability* between states x and y . In other words, s_{xy} is the fraction of the frequency of state x that is the result of transitions from state y to state x .

Corollary The 12 off-diagonal entries of the rate matrix, Q (note the off-diagonal entries determine the diagonal entries, since the rows of Q sum to zero) can be completely determined by 9 numbers; these are: 6 exchangeability terms and 3 stationary frequencies π_x , (since the stationary frequencies sum to 1).

Scaling of branch lengths

By comparing extant sequences, one can determine the amount of sequence divergence. This raw measurement of divergence provides information about the number of changes that have occurred along the path separating the sequences. The simple count of differences (the Hamming distance) between sequences will often underestimate the number of substitution because of multiple hits (see homoplasy). Trying to estimate the exact number of changes that have occurred is difficult, and usually not necessary. Instead, branch lengths (and path lengths) in phylogenetic analyses are usually expressed in the expected number of changes per site. The path length is the product of the duration of the path in time and the mean rate of substitutions. While their product can be estimated, the rate and time are not identifiable from sequence divergence.

The descriptions of rate matrices on this page accurately reflect the relative magnitude of different substitutions, but these rate matrices are **not** scaled such that a branch length of 1 yields one expected change. This scaling can be accomplished by multiplying every element of the matrix by the same factor, or simply by scaling the branch lengths. If we use the β to denote the scaling factor, and ν to denote the branch length measured in the expected number of substitutions per site then βν is used in the transition probability formulae below in place of μ*t*. Note that ν is a parameter to be estimated from data, and is referred to as the branch length, while β is simply a number that can be calculated from the rate matrix (it is not a separate free parameter).

The value of β can be found by forcing the expected rate of flux of states to 1. The diagonal entries of the rate-matrix (the *Q* matrix) represent -1 times the rate of leaving each state. For time-reversible models, we know the equilibrium state frequencies (these are simply the π_i parameter value for state *i*). Thus we can find the expected rate of change by calculating the sum of flux out of each state weighted by the proportion of sites that are expected to be in that class. Setting β to be the reciprocal of this sum will guarantee that scaled process has an expected flux of 1:

$$\beta = 1/\left(-\sum_i \pi_i \mu_{ii}\right)$$

For example, in the Jukes-Cantor, the scaling factor would be *4/(3μ)* because the rate of leaving each state is *3μ/4*.

Most common models of DNA evolution

JC69 model (Jukes and Cantor 1969)

JC69, the Jukes and Cantor 1969 model,[1] is the simplest substitution model. There are several assumptions. It assumes equal base frequencies $\left(\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}\right)$ and equal mutation rates. The only parameter of this model is therefore μ, the overall substitution rate. As previously mentioned, this variable becomes a constant when we normalize the mean-rate to 1.

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

$$P = \begin{pmatrix} \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} \end{pmatrix}$$

When branch length, ν, is measured in the expected number of changes per site then:

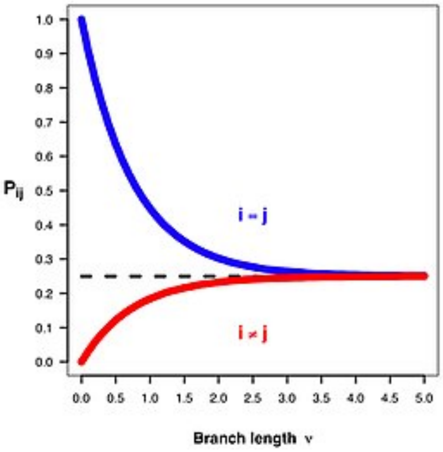
$$P_{ij}(\nu) = \begin{cases} \frac{1}{4}+\frac{3}{4}e^{-4\nu/3} & \text{if } i = j \\ \frac{1}{4}-\frac{1}{4}e^{-4\nu/3} & \text{if } i \neq j \end{cases}$$

It is worth noticing that $\nu = \frac{3}{4}t\mu = (\frac{\mu}{4}+\frac{\mu}{4}+\frac{\mu}{4})t$ what stands for sum of any column (or row) of matrix Q multiplied by time and thus means expected number of substitutions in time t (branch duration) for each particular site (per site) when the rate of substitution equals μ.

Given the proportion p of sites that differ between the two sequences the Jukes-Cantor estimate of the evolutionary distance (in terms of the expected number of changes) between two sequences is given by

$$\hat{d} = -\frac{3}{4}\ln(1 - \frac{4}{3}p) = \hat{\nu}$$

The p in this formula is frequently referred to as the p-distance. It is a sufficient statistic for calculating the Jukes-Cantor distance correction, but is not sufficient for the calculation of the evolutionary distance under the more complex models that follow (also note that p used in subsequent formulae is not identical to the "p-distance").



Probability P_{ij} of changing from initial state i to final state j as a function of the branch length (ν) for JC69. Red curve: nucleotide states i and j are different. Blue curve: initial and final states are the same. After a long time, probabilities tend to the nucleotide equilibrium frequencies (0.25: dashed line).

K80 model (Kimura 1980)

K80, the Kimura 1980 model,[2] often referred to as **Kimura's two parameter model** (or the **K2P model**), distinguishes between transitions ($A \leftrightarrow G$, i.e. from purine to purine, or $C \leftrightarrow T$, i.e. from pyrimidine to pyrimidine) and transversions (from purine to pyrimidine or vice versa). In Kimura's original description of the model the α and β were used to denote the rates of these types of substitutions, but it is now more common to set the rate of transversions to 1 and use κ to denote the transition/transversion rate ratio (as is done below). The K80 model assumes that all of the bases are equally frequent ($\pi_A = \pi_G = \pi_C = \pi_T = 0.25$).

Rate matrix $Q = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$ with columns corresponding to A, G, C, and T, respectively.

The Kimura two-parameter distance is given by:

$$K = -\frac{1}{2}\ln((1 - 2p - q)\sqrt{1 - 2q})$$

where p is the proportion of sites that show transitional differences and q is the proportion of sites that show transversional differences.

K81 model (Kimura 1981)

K81, the Kimura 1981 model,[3] often called **Kimura's three parameter model** (K3P model) or the Kimura three substitution type (K3ST) model, has distinct rates for transitions and two distinct types of transversions. The two transversion types are those that conserve the weak/strong properties of the nucleotides (i.e., $A \leftrightarrow T$ and $C \leftrightarrow G$, denoted by symbol γ [3]) and those that conserve the amino/keto properties of the nucleotides (i.e., $A \leftrightarrow C$ and $G \leftrightarrow T$, denoted by symbol β [3]). The K81 model assumes that all equilibrium base frequencies are equal (i.e., $\pi_A = \pi_G = \pi_C = \pi_T = 0.25$).

Rate matrix $Q = \begin{pmatrix} * & \alpha & \beta & \gamma \\ \alpha & * & \gamma & \beta \\ \beta & \gamma & * & \alpha \\ \gamma & \beta & \alpha & * \end{pmatrix}$ with columns corresponding to A, G, C, and T, respectively.

The K81 model is used much less often than the K80 (K2P) model for distance estimation and it is seldom the best-fitting model in maximum likelihood phylogenetics. Despite these facts, the K81 model has continued to be studied in the context of mathematical phylogenetics.[4][5][6] One important property is the ability to perform a Hadamard transform assuming the site patterns were generated on a tree with nucleotides evolving under the K81 model.[7][8][9]

When used in the context of phylogenetics the Hadamard transform provides an elegant and fully invertible means to calculate expected site pattern frequencies given a set of branch lengths (or vice versa). Unlike many maximum likelihood calculations, the relative values for α, β, and γ can vary across branches and the Hadamard transform can even provide evidence that the data do not fit a tree. The Hadamard transform can also be combined with a wide variety of methods to accommodate among-sites rate heterogeneity,[10] using continuous distributions rather than the discrete approximations typically used in maximum likelihood phylogenetics[11] (although one must sacrifice the invertibility of the Hadamard transform to use certain among-sites rate heterogeneity distributions[10]).

F81 model (Felsenstein 1981)

F81, the Felsenstein's 1981 model,[12] is an extension of the JC69 model in which base frequencies are allowed to vary from 0.25 ($\pi_A \neq \pi_G \neq \pi_C \neq \pi_T$)

Rate matrix:

$$Q = \begin{pmatrix} * & \pi_G & \pi_C & \pi_T \\ \pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \pi_T \\ \pi_A & \pi_G & \pi_C & * \end{pmatrix}$$

When branch length, ν, is measured in the expected number of changes per site then:

$$\beta = 1/(1 - \pi_A^2 - \pi_C^2 - \pi_G^2 - \pi_T^2)$$
$$P_{ij}(\nu) = \begin{cases} e^{-\beta\nu} + \pi_j \left(1 - e^{-\beta\nu}\right) & \text{if } i = j \\ \pi_j \left(1 - e^{-\beta\nu}\right) & \text{if } i \neq j \end{cases}$$

HKY85 model (Hasegawa, Kishino and Yano 1985)

HKY85, the Hasegawa, Kishino and Yano 1985 model,[13] can be thought of as combining the extensions made in the Kimura80 and Felsenstein81 models. Namely, it distinguishes between the rate of transitions and transversions (using the κ parameter), and it allows unequal base frequencies ($\pi_A \neq \pi_G \neq \pi_C \neq \pi_T$). [Felsenstein described a similar (but not equivalent) model in 1984 using a different parameterization;[14] that latter model is referred to as the F84 model.[15]]

Rate matrix $Q = \begin{pmatrix} * & \kappa\pi_G & \pi_C & \pi_T \\ \kappa\pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \kappa\pi_T \\ \pi_A & \pi_G & \kappa\pi_C & * \end{pmatrix}$

If we express the branch length, ν in terms of the expected number of changes per site then:

$$\beta = \frac{1}{2(\pi_A + \pi_G)(\pi_C + \pi_T) + 2\kappa[(\pi_A\pi_G) + (\pi_C\pi_T)]}$$
$$P_{AA}(\nu, \kappa, \pi) = \left[\pi_A \left(\pi_A + \pi_G + (\pi_C + \pi_T)e^{-\beta\nu}\right) + \pi_G e^{-(1+(\pi_A+\pi_G)(\kappa-1.0))\beta\nu}\right] / (\pi_A + \pi_G)$$
$$P_{AC}(\nu, \kappa, \pi) = \pi_C \left(1.0 - e^{-\beta\nu}\right)$$

$$P_{AG}(\nu,\kappa,\pi) = \left[\pi_G\left(\pi_A+\pi_G+(\pi_C+\pi_T)e^{-\beta\nu}\right) - \pi_G e^{-(1+(\pi_A+\pi_G)(\kappa-1.0))\beta\nu}\right] / (\pi_A+\pi_G)$$
$$P_{AT}(\nu,\kappa,\pi) = \pi_T\left(1.0 - e^{-\beta\nu}\right)$$

and formula for the other combinations of states can be obtained by substituting in the appropriate base frequencies.

T92 model (Tamura 1992)

T92, the Tamura 1992 model,[16] is a mathematical method developed to estimate the number of nucleotide substitutions per site between two DNA sequences, by extending Kimura's (1980) two-parameter method to the case where a G+C content bias exists. This method will be useful when there are strong transition-transversion and G+C-content biases, as in the case of *Drosophila* mitochondrial DNA.[16]

T92 involves a single, compound base frequency parameter $\theta \in (0,1)$ (also noted $\pi_{GC}) = \pi_G + \pi_C = 1 - (\pi_A + \pi_T)$

As T92 echoes the Chargaff's second parity rule — pairing nucleotides do have the same frequency on a single DNA strand, G and C on the one hand, and A and T on the other hand — it follows that the four base frequences can be expressed as a function of π_{GC}

$\pi_G = \pi_C = \dfrac{\pi_{GC}}{2}$ and $\pi_A = \pi_T = \dfrac{(1-\pi_{GC})}{2}$

Rate matrix $Q = \begin{pmatrix} * & \kappa\pi_{GC}/2 & \pi_{GC}/2 & (1-\pi_{GC})/2 \\ \kappa(1-\pi_{GC})/2 & * & \pi_{GC}/2 & (1-\pi_{GC})/2 \\ (1-\pi_{GC})/2 & \pi_{GC}/2 & * & \kappa(1-\pi_{GC})/2 \\ (1-\pi_{GC})/2 & \pi_{GC}/2 & \kappa\pi_{GC}/2 & * \end{pmatrix}$

The evolutionary distance between two DNA sequences according to this model is given by

$$d = -h\ln(1-\frac{p}{h}-q) - \frac{1}{2}(1-h)\ln(1-2q)$$

where $h = 2\theta(1-\theta)$ and θ is the G+C content ($\pi_{GC} = \pi_G + \pi_C$).

TN93 model (Tamura and Nei 1993)

TN93, the Tamura and Nei 1993 model,[17] distinguishes between the two different types of transition; i.e. ($A \leftrightarrow G$) is allowed to have a different rate to ($C \leftrightarrow T$). Transversions are all assumed to occur at the same rate, but that rate is allowed to be different from both of the rates for transitions.

TN93 also allows unequal base frequencies ($\pi_A \neq \pi_G \neq \pi_C \neq \pi_T$).

Rate matrix $Q = \begin{pmatrix} * & \kappa_1\pi_G & \pi_C & \pi_T \\ \kappa_1\pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \kappa_2\pi_T \\ \pi_A & \pi_G & \kappa_2\pi_C & * \end{pmatrix}$

GTR model (Tavaré 1986)

GTR, the Generalised time-reversible model of Tavaré 1986,[18] is the most general neutral, independent, finite-sites, time-reversible model possible. It was first described in a general form by Simon Tavaré in 1986.[18]

GTR parameters consist of an equilibrium base frequency vector, $\Pi = (\pi_A, \pi_G, \pi_C, \pi_T)$, giving the frequency at which each base occurs at each site, and the rate matrix

$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & -(\gamma\pi_A + \epsilon\pi_G + \eta\pi_C) \end{pmatrix}$$

Where

$$\alpha = r(A \to G) = r(G \to A)$$
$$\beta = r(A \to C) = r(C \to A)$$
$$\gamma = r(A \to T) = r(T \to A)$$
$$\delta = r(G \to C) = r(C \to G)$$
$$\epsilon = r(G \to T) = r(T \to G)$$
$$\eta = r(C \to T) = r(T \to C)$$

are the transition rate parameters.

Therefore, GTR (for four characters, as is often the case in phylogenetics) requires 6 substitution rate parameters, as well as 4 equilibrium base frequency parameters. However, this is usually eliminated down to 9 parameters plus μ, the overall number of substitutions per unit time. When measuring time in substitutions (μ=1) only 8 free parameters remain.

In general, to compute the number of parameters, one must count the number of entries above the diagonal in the matrix, i.e. for n trait values per site $\frac{n^2 - n}{2}$, and then add n for the equilibrium base frequencies, and subtract 1 because μ is fixed. One gets

$$\frac{n^2 - n}{2} + n - 1 = \frac{1}{2}n^2 + \frac{1}{2}n - 1.$$

For example, for an amino acid sequence (there are 20 "standard" amino acids that make up proteins), one would find there are 209 parameters. However, when studying coding regions of the genome, it is more common to work with a codon substitution model (a codon is three bases and codes for one amino acid in a protein). There are $4^3 = 64$ codons, but the rates for transitions between codons which differ by more than one base is assumed to be zero. Hence, there are $\frac{20 \times 19 \times 3}{2} + 64 - 1 = 633$ parameters.

See also

- Molecular evolution
- Molecular clock
- UPGMA

References

1. Jukes TH, Cantor CR (1969). *Evolution of Protein Molecules*. New York: Academic Press. pp. 21–132.

2. Kimura M (December 1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences". *Journal of Molecular Evolution*. **16** (2): 111–20. Bibcode:1980JMolE..16..111K (https://ui.adsabs.harvard.edu/abs/1980JMolE..16..111K). doi:10.1007/BF01731581 (https://doi.org/10.1007%2FBF01731581). PMID 7463489 (https://pubmed.ncbi.nlm.nih.gov/7463489).
3. Kimura M (January 1981). "Estimation of evolutionary distances between homologous nucleotide sequences" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC319072). *Proceedings of the National Academy of Sciences of the United States of America*. **78** (1): 454–8. Bibcode:1981PNAS...78..454K (https://ui.adsabs.harvard.edu/abs/1981PNAS...78..454K). doi:10.1073/pnas.78.1.454 (https://doi.org/10.1073%2Fpnas.78.1.454). PMC 319072 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC319072). PMID 6165991 (https://pubmed.ncbi.nlm.nih.gov/6165991).
4. Bashford JD, Jarvis PD, Sumner JG, Steel MA (2004-02-25). "U (1) × U (1) × U (1) symmetry of the Kimura 3ST model and phylogenetic branching processes". *Journal of Physics A: Mathematical and General*. **37** (8): L81–L89. arXiv:q-bio/0310037 (https://arxiv.org/abs/q-bio/0310037). doi:10.1088/0305-4470/37/8/L01 (https://doi.org/10.1088%2F0305-4470%2F37%2F8%2FL01).
5. Sumner JG, Charleston MA, Jermiin LS, Jarvis PD (August 2008). "Markov invariants, plethysms, and phylogenetics". *Journal of Theoretical Biology*. **253** (3): 601–15. doi:10.1016/j.jtbi.2008.04.001 (https://doi.org/10.1016%2Fj.jtbi.2008.04.001). PMID 18513747 (https://pubmed.ncbi.nlm.nih.gov/18513747).
6. Sumner JG, Jarvis PD, Holland BR (December 2014). "A tensorial approach to the inversion of group-based phylogenetic models" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4268818). *BMC Evolutionary Biology*. **14** (1): 236. doi:10.1186/s12862-014-0236-6 (https://doi.org/10.1186%2Fs12862-014-0236-6). PMC 4268818 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4268818). PMID 25472897 (https://pubmed.ncbi.nlm.nih.gov/25472897).
7. Hendy MD, Penny D, Steel MA (April 1994). "A discrete Fourier analysis for evolutionary trees" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC43572). *Proceedings of the National Academy of Sciences of the United States of America*. **91** (8): 3339–43. Bibcode:1994PNAS...91.3339H (https://ui.adsabs.harvard.edu/abs/1994PNAS...91.3339H). doi:10.1073/pnas.91.8.3339 (https://doi.org/10.1073%2Fpnas.91.8.3339). PMC 43572 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC43572). PMID 8159749 (https://pubmed.ncbi.nlm.nih.gov/8159749).
8. Hendy MD (2005). "Hadamard conjugation: an analytic tool for phylogenetics" (https://books.google.com/books?hl=en&lr=&id=Ao_MNOaFRHEC&oi=fnd&pg=PA143). In Gascuel O (ed.). *Mathematics of Evolution and Phylogeny*. Oxford University Press. pp. 143–177. ISBN 978-0198566106.
9. Hendy MD, Snir S (July 2008). "Hadamard conjugation for the Kimura 3ST model: combinatorial proof using path sets". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. **5** (3): 461–71. doi:10.1109/TCBB.2007.70227 (https://doi.org/10.1109%2FTCBB.2007.70227). PMID 18670048 (https://pubmed.ncbi.nlm.nih.gov/18670048).
10. Waddell PJ, Penny D, Moore T (August 1997). "Hadamard conjugations and modeling sequence evolution with unequal rates across sites". *Molecular Phylogenetics and Evolution*. **8** (1): 33–50. doi:10.1006/mpev.1997.0405 (https://doi.org/10.1006%2Fmpev.1997.0405). PMID 9242594 (https://pubmed.ncbi.nlm.nih.gov/9242594).
11. Yang Z (September 1994). "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods". *Journal of Molecular Evolution*. **39** (3): 306–14. Bibcode:1994JMolE..39..306Y (https://ui.adsabs.harvard.edu/abs/1994JMolE..39..306Y). CiteSeerX 10.1.1.305.951 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.305.951). doi:10.1007/BF00160154 (https://doi.org/10.1007%2FBF00160154). PMID 7932792 (https://pubmed.ncbi.nlm.nih.gov/7932792).

12. Felsenstein J (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach". *Journal of Molecular Evolution*. **17** (6): 368–76. Bibcode:1981JMolE..17..368F (https://ui.adsabs.harvard.edu/abs/1981JMolE..17..368F). doi:10.1007/BF01734359 (https://doi.org/10.1007%2FBF01734359). PMID 7288891 (https://pubmed.ncbi.nlm.nih.gov/7288891).
13. Hasegawa M, Kishino H, Yano T (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA". *Journal of Molecular Evolution*. **22** (2): 160–74. Bibcode:1985JMolE..22..160H (https://ui.adsabs.harvard.edu/abs/1985JMolE..22..160H). doi:10.1007/BF02101694 (https://doi.org/10.1007%2FBF02101694). PMID 3934395 (https://pubmed.ncbi.nlm.nih.gov/3934395).
14. Kishino H, Hasegawa M (August 1989). "Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea". *Journal of Molecular Evolution*. **29** (2): 170–9. Bibcode:1989JMolE..29..170K (https://ui.adsabs.harvard.edu/abs/1989JMolE..29..170K). doi:10.1007/BF02100115 (https://doi.org/10.1007%2FBF02100115). PMID 2509717 (https://pubmed.ncbi.nlm.nih.gov/2509717).
15. Felsenstein J, Churchill GA (January 1996). "A Hidden Markov Model approach to variation among sites in rate of evolution" (https://doi.org/10.1093/oxfordjournals.molbev.a025575). *Molecular Biology and Evolution*. **13** (1): 93–104. doi:10.1093/oxfordjournals.molbev.a025575 (https://doi.org/10.1093%2Foxfordjournals.molbev.a025575). PMID 8583911 (https://pubmed.ncbi.nlm.nih.gov/8583911).
16. Tamura K (July 1992). "Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases" (https://doi.org/10.1093/oxfordjournals.molbev.a040752). *Molecular Biology and Evolution*. **9** (4): 678–87. doi:10.1093/oxfordjournals.molbev.a040752 (https://doi.org/10.1093%2Foxfordjournals.molbev.a040752). PMID 1630306 (https://pubmed.ncbi.nlm.nih.gov/1630306).
17. Tamura K, Nei M (May 1993). "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees" (https://doi.org/10.1093/oxfordjournals.molbev.a040023). *Molecular Biology and Evolution*. **10** (3): 512–26. doi:10.1093/oxfordjournals.molbev.a040023 (https://doi.org/10.1093%2Foxfordjournals.molbev.a040023). PMID 8336541 (https://pubmed.ncbi.nlm.nih.gov/8336541).
18. Tavaré S (1986). "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences" (http://www.damtp.cam.ac.uk/user/st321/CV_&_Publications_files/STpapers-pdf/T86.pdf) (PDF). *Lectures on Mathematics in the Life Sciences*. **17**: 57–86.

Further reading

- Gu X, Li WH (September 1992). "Higher rates of amino acid substitution in rodents than in humans". *Molecular Phylogenetics and Evolution*. **1** (3): 211–4. doi:10.1016/1055-7903(92)90017-B (https://doi.org/10.1016%2F1055-7903%2892%2990017-B). PMID 1342937 (https://pubmed.ncbi.nlm.nih.gov/1342937).
- Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D (February 1996). "Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis". *Molecular Phylogenetics and Evolution*. **5** (1): 182–7. doi:10.1006/mpev.1996.0012 (https://doi.org/10.1006%2Fmpev.1996.0012). PMID 8673286 (https://pubmed.ncbi.nlm.nih.gov/8673286).

External links

- DAWG: DNA Assembly With Gaps (http://scit.us/projects/dawg) — free software for simulating sequence evolution

Retrieved from "https://en.wikipedia.org/w/index.php?title=Models_of_DNA_evolution&oldid=969627326"

This page was last edited on 26 July 2020, at 15:36 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.