ISSN: 2378-3648

Xia. J Genet Genome Res 2017, 4:031
DOI: 10.23937/2378-3648/1410031
Volume 4 | Issue 1
Open Access

Journal of
Genetics and Genome Research

REVIEW ARTICLE

Deriving Transition Probabilities and Evolutionary Distances from Substitution Rate Matrix by Probability Reasoning

Xuhua Xia*[1,2]

[1]*Department of Biology, University of Ottawa, Canada*

[2]*Ottawa Institute of Systems Biology, Ottawa, Canada*

***Corresponding author:** *Xuhua Xia, Department of Biology, Ottawa Institute of Systems Biology, University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A, Ottawa, Ontario, K1N 6N5, Canada, Tel: (613)-562-5800, Ext: 6886, Fax: (613)-562-5486, E-mail: xxia@uottawa.ca*

Abstract

Substitution rate matrices are used to correct multiple hits at the same sites, which requires the derivation of transition probabilities and evolutionary distances from substitution rate matrices. The derivation is essential in molecular phylogenetics and phylogenomics, and represents the only statistically sound way for developing scoring matrices used in sequence alignment and local string matching (e.g., BLAST and FASTA). Three different approaches are frequently used for deriving transition probabilities and evolutionary distances: 1) The probability reasoning, 2) Solving partial differential equations, and 3) Matrix exponential and logarithm. The first approach demands the least amount of mathematical skills but offers the best way for conceptual understanding, and can often generate nice mathematical expressions of transition probabilities and evolutionary distances. This review represents the most systematic and comprehensive numerical illustration of the first approach.

Keywords

Substitution model, Substitution rate, Transition probability, Evolutionary distance

Introduction

Substitutions occur over time and can overwrite each other at the same nucleotide or amino acid site. When we compare two homologous nucleotide sequences and find differences in N sites, the actual number of substitutions (designated by M) could be much greater than N because multiple substitutions could have happened at the same site, overwriting each other. Substitution models are used to infer the observable M from the observed substitutions from sequence comparisons.

Many substitution models have been proposed for nucleotide, amino acid and codon sequences. All substitution models used in molecular phylogenetics are Markov chain models characterized by 1) Either a transition probability matrix (P) with discrete time or a rate matrix (Q) in continuous time where P can be derived from Q, and 2) Equilibrium frequencies. The general form of an instantaneous rate matrix for nucleotide sequences is, in the order of A, G, C, and T:

$$Q = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{bmatrix} - & a & b & c \\ g & - & d & e \\ h & i & - & f \\ j & k & l & - \end{bmatrix} \end{array} \quad (1)$$

Transition probability matrix, often referred to as the P matrix, specifies the probability of a nucleotide or amino acid changing into another one after time t. It is needed to calculate likelihood and to derive evolutionary distances, and consequently is needed phylogenetics based on the maximum likelihood and distance-based methods as well as Bayesian inference. Whether a substitution model can be implemented for phylogenetic analysis essentially depends on whether the model's transition probabilities can be calculated.

There are three ways to obtain transition probabilities from the Q matrix [1,2] :1) By probability reasoning, 2) By solving differential equations involving rates, and 3) By taking the matrix exponential of the rate matrix. The last two require some mathematical background in calculus and linear algebra. The first, in contrast, demands little mathematical skill except for careful book-keeping and solving simultaneous equations. This approach is particularly relevant to biological students not only for gaining a concep-

Citation: Xia X (2017) Deriving Transition Probabilities and Evolutionary Distances from Substitution Rate Matrix by Probability Reasoning. J Genet Genome Res 3:031. doi.org/10.23937/2378-3648/1410031

Received: April 03, 2017: **Accepted:** July 24, 2017: **Published:** July 27, 2017

tual understanding of the substitution models, but also to deriving nice mathematical expressions for transition probabilities and evolutionary distances. New researchers often ask why we can derive evolutionary distances between two aligned sequences for the TN93 model [3] but cannot for the simpler HKY85 model [4] which is a special case of the TN93 model, yet another model, F84 (used in PHYLIP since 1984), which is also a special case of the TN93 model, can have its evolutionary distance readily derived. One can easily obtain answers to such questions by taking the first approach. However, the first approach is not of general purpose and cannot handle very complicated substitution models. In contrast, the last one can be used with any substitution models specified by a rate matrix from which the matrix exponential can be obtained. In short, all these approaches need to be learned by anyone wishing to become a molecular phylogeneticist, but this paper will focus only on the probability reasoning approach illustrated with JC69 [5], K80 [6], F84 (the model used in PHYLIP since 1984), HKY85 [4], and TN93 [3] models.

Probability Reasoning to Obtain Transition Probabilities and Evolutionary Distances

Felsenstein [1] presented nice examples of probability reasoning to derive transition probabilities and evolutionary distances from rate matrices. This section presents the approach in a more systematic and accessible way.

JC69 model

Consider nucleotide A in the JC69 model (Figure 1a). Imagine that the nucleotide has a rate α of changing into any of the four nucleotides, i.e., including changing to itself (Figure 1b). This is effectively the same specification as the JC69 model. After time t, the expected number of substitutions is 4αt and the probability of no substitution is $p(x = 0, \alpha, t) = e^{-4\alpha t}$ according to the Poisson distribution, and the probability of having at least one change is then $p(x \geq 1, \alpha, t) = 1-e^{-4\alpha t}$ (Figure 1c). Because nucleotide A can change into any one of the four nucleotides (including nucleotide A itself), each nucleotide gets 1/4 of $p(x \geq 1, \alpha, t)$. We therefore have in Figure 1.

$$p_{ij}(t) = \frac{p(x \geq 1, \alpha, t)}{4} = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \quad (2)$$

The transition probability $p_{ii}(t)$ is the summation of two probabilities: the probability of no change (which is $e^{-4\alpha t}$) and the probability of changing to itself which is the same as specified in Eq. (2), as shown in Figure 1e, i.e.,

$$p_{ii}(t) = e^{-4\alpha t} + \frac{p(x \geq 1, \alpha, t)}{4} = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \quad (3)$$

The transition probability matrix for the JC69 model has only two distinct elements. The four diagonal elements are the same as specified in Eq. (3) and all the off-diagonal elements are the same as specified in Eq. (2). Each row in P adds up to 1 as a nucleotide can either stay the same or change into some other nucleotides.

There are some quick ways to check the derived transition probabilities. First, we note that when t approaches infinity, then all entries in matrix P approaches ¼ if α > 0. This is what we have expected. Second, when t = 0, then all diagonal elements in matrix P are 1 and all off-diagonal elements are zero. This is again what we expected. Third, if α is zero, then no change is possible, and we again expect all diagonal elements in matrix P to be 1 and all off-diagonal elements to be zero, which is also true.

From p_{ij} in Eq. (2), the expected proportion of sites that are different between two aligned homologous sequences (p_{diff}) is $3^{*}p_{ij}(t)$, i.e.,

$$p_{diff} = 3p_{ij}(t) = \frac{3}{4} - \frac{3}{4} e^{-4\alpha t} \quad (4)$$

Note that p_{diff} approaches ¾ when t is infinitely large, which means that multiple substitutions can no longer be corrected. Eq. (4) offers another way of deriving $p_{ii}(t)$ in Eq. (3), i.e., $p_{ii}(t)$ is simply $1-p_{diff}$.

Eq. (4) allows us to derive the JC69 distance (D_{JC69}) because a distance is defined as μt where μ is the substitution rate which is equal to 3α in the JC69 model. This is the same as the distance that you have driven is the product of the speed (rate) and time. Given that $D_{JC69} = 3\alpha t$, we can derive D_{JC69} (Figure 1g) by substituting $\alpha t = D_{JC69}/3$ into Eq. (4), i.e.,

$$D_{JC69} = -\frac{3}{4} \ln\left(1 - \frac{4p_{diff}}{3}\right) \quad (5)$$

Where p_{diff} (the expected number of sites that are different between the two homologous sequences) can be approximated by the observed proportion of sites ($p_{diff.obs}$) differing between the two aligned sequences. Note that $p_{diff.obs}$ may differ from p_{diff} even when the underlying substitution model indeed follows JC69 because of 1) Stochastic factors due to limited aligned length of the two sequences, and 2) Distortion caused by suboptimal sequence alignment. Thus, although p_{diff} in Eq. (4) cannot be greater than 0.75, $p_{diff.obs}$ could, even when sequences evolve strictly according to the JC69 model. D_{JC69} is not defined when $p_{diff} \geq 0.75$ as there is no logarithm for 0 or negative values.

We can optionally show that D_{JC69} in Eq. (5) is a maximum likelihood distance. For two aligned sequences of length N, designate the number of sites that differ between the two sequences as N_D and the number of sites identical between the two sites as $(N-N_D)$. Now the likelihood function is:

$$L = \left(\frac{1}{4}\right)^{N} p_{ii}^{(N-N_D)} (1-p_{ii})^{N_D}$$

$$\ln L = N \ln\left(\frac{1}{4}\right) + (N-N_D)\ln(p_{ii}) + N_D \ln(1-p_{ii})$$

$$= N \ln\left(\frac{1}{4}\right) + (N-N_D)\ln\left(\frac{1}{4} + \frac{3}{4} e^{-4D_{JC69}/3}\right) + N_D \ln\left(\frac{3}{4} - \frac{3}{4} e^{-4D_{JC69}/3}\right) \quad (6)$$

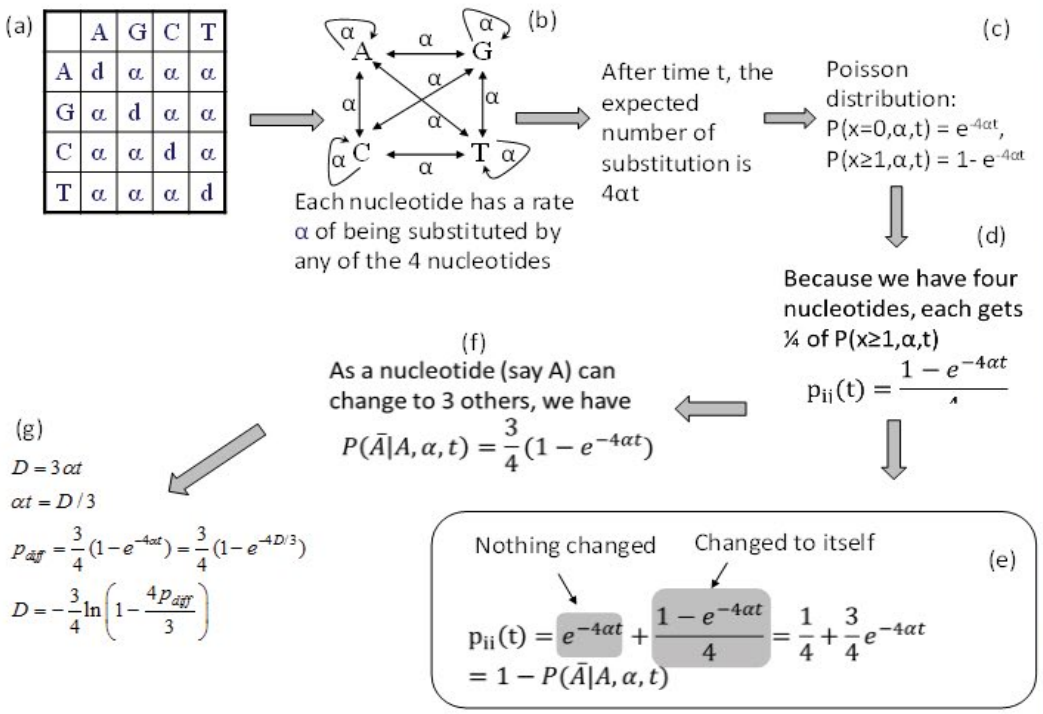

Figure 1: Derivation of transition probabilities and the evolutionary distance (D) based on the JC69 model. The d value in the diagonal of the rate matrix (a) is constrained by the row sum equal to 0, i.e., d = -3α. P(j|i,t) means the probability of changing from the original nucleotide i to nucleotide j after time t, and is synonymous to $p_{ij}(t)$ or simply p_{ij} in this paper.

```
S1: AAG CCT CGG GGC CCT TAT TTT TTG
    ||  |   ||| ||| |   ||| ||| ||
S2: AAT CTC CGG GGC CTC TAT TTT TTT
```

Figure 2: Two homologous sequences for illustrating computation of pairwise evolutionary distances.

Where the constant term N*ln(1/4) can be dropped in maximizing lnL to obtain the distance estimate, but needs to be kept when performing likelihood ratio test for comparing different substitution models (e.g., JC69 against TN93).

We take the derivative of lnL with respect to D_{JC69}, set the derivative to 0 and solve for D_{JC69}. The resulting D_{JC69} is exactly the same as that in Eq. (5). I used D instead of D_{JC69} in the equations below:

$$\frac{d\ln L}{dD} = -\frac{(N-N_D)e^{-4D/3}}{\frac{1}{4}+\frac{3}{4}e^{-4D/3}} + \frac{N_D e^{-4D/3}}{\frac{3}{4}-\frac{3}{4}e^{-4D/3}} = 0$$

$$D = -\frac{3}{4}\ln\left(\frac{3N-4N_D}{3N}\right) = -\frac{3}{4}\ln\left(1-\frac{4p_{diff}}{3}\right) \quad (7)$$

The variance of D_{JC69} (designated as V_{JC69}) is obtained as the negative reciprocal of the second derivative of lnL:

$$V_{JC69} = -\frac{1}{\frac{d^2\ln L}{dD_{JC69}^2}} = \frac{p_{diff}(1-p_{diff})}{L\left(1-\frac{4p_{diff}}{3}\right)^2} \quad (8)$$

Note that V_{JC69} decreases with sequence length L as one would have expected. We illustrate the application of Eqs. (5) and (8) by using the aligned sequences in Figure 2 where N = 24, N_D = 6, and p_{diff} = 6/24 = 0.25. So D_{JC69} = 0.3041, and var (D_{JC69}) = 0.0176.

The equilibrium frequencies of the π vector can be derived by set t = ∞ in Eqs. (2) and (3) which leads to $p_{ii} = p_{ij}$ = ¼. This implies that equilibrium frequencies of the four nucleotides will be equal for the JC69 model. This is not surprising because the frequencies did not even appear in the rate matrix (Figure 1a).

K80 model

The K80 model has a transition substitution rate α and a transversion rate β (Figure 3a). We will focus on nucleotide A and conceptualize the model with two events (Figure 3b), in contrast to only one event in the JC69 model. The first event (e_1) occurs when nucleotide A changes into any of the four nucleotides (including to itself). In other words, the original A is replace by a nucleotide randomly drawn from a nucleotide pool with equal nucleotide frequencies. This event occurs with a rate β. The second event (e_2) occurs when nucleotide A changes either to G or to itself, i.e., the original A is replace by a nucleotide randomly drawn from a purine pool with equal number of A and G. This e_2 occurs with a rate γ. Thus the transition rate α equals β+γ according to this conceptualization. Note that, whenever e_1 happens, the original nucleotide is replace by any one of the four nucleotides with equal probability, no matter how many e_2 events has occur before or after the occurrence of e_1. It might help to think of a long sequence with L sites

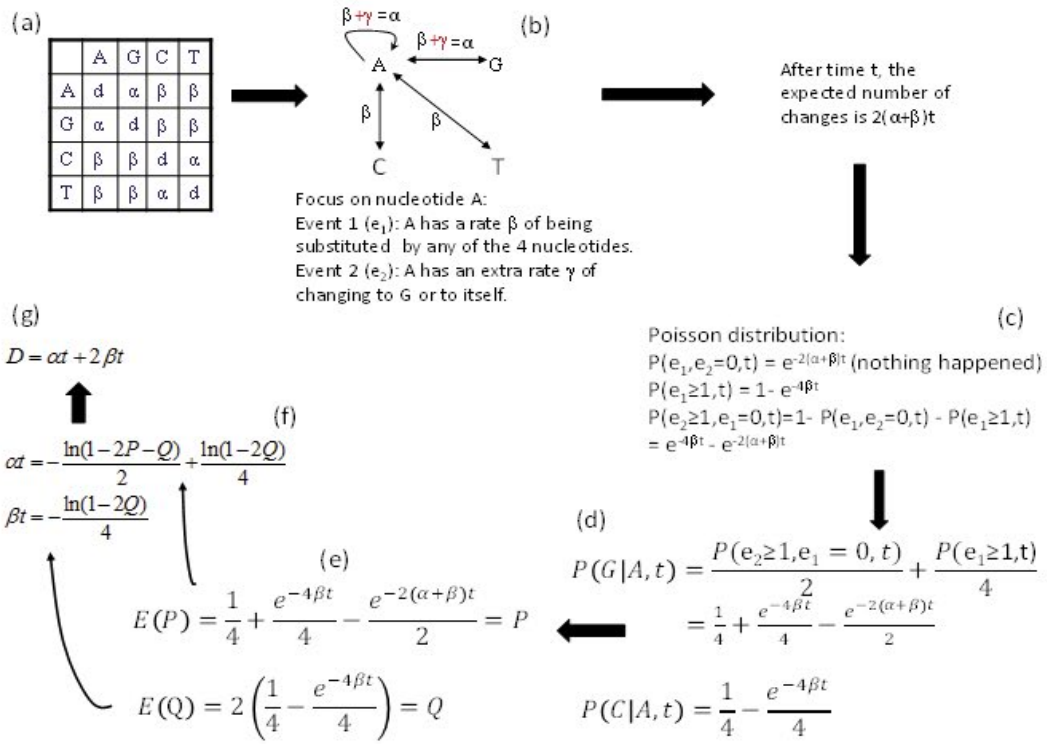

Figure 3: Derivation of transition probabilities and the evolutionary distance (D) based on the K80 model. The rate matrix (a) has the diagonal elements (d) constrained by the row sum equal to 0, i.e., d = -α - 2β. P and Q are the observed proportion of transitional and transversional changes between two aligned homologous sequences. Equating them to their respective expected values, E(P) and E(Q), leads to the solution of αt and βt shown, and the evolutionary distance D. P(j|i,t) means the probability of changing from the original nucleotide i to nucleotide j after time t, and is synonymous to $p_{ij}(t)$ or simply p_{ij} in this paper.

being A at time 0. If these L sites each have experienced at least one e_2 event, then these sites will either be A or G with equal probability (i.e., 0.5), and we expect to have L/2 sites being A and the other L/2 sites being G. In contrast, if each of these L sites has experienced at least one e_1 event, then the site will be replaced by either A, C, G, or T with equal probability, and we expect to observe A, C, G and T in L/4 sites each. Any e_2 events occurring before or after the e_1 event do not change this expectation. This means that e_1 erases e_2, but not vice versa. The probability that an e_2 event happened is informative only when no e_1 event has happened.

After time t, the expected number of substitutions is 2(α+β)t, i.e., the nucleotide A has two ways of change with a rate of α (to A and to G) and another two ways of change with a rate β (to C and to T), so the probability of no change, according to Poisson distribution, is

$$p(e_1 = 0, e_2 = 0, t) = e^{-2(\alpha+\beta)t} \quad (9)$$

Note that α is conceptualized as (β+γ) in Figure 3, so $e^{-2(\alpha+\beta)t}$ in Eq. (9) is equivalent to $e^{-(4\beta+2\gamma)t}$. The probability that at least one e_1 event has occurred is

$$p(e_1 > 0, t) = 1 - e^{-4\beta t} \quad (10)$$

Thus, the probability that at least one e_2 has occurred but e_1 has not occurred is simply

$$\begin{aligned} p(e_2 > 0, e_1 = 0, t) &= 1 - p(e_1 = 0, e_2 = 0, t) - p(e_1 > 0, t) \\ &= 1 - e^{-2(\alpha+\beta)t} - (1 - e^{-4\beta t}) \\ &= e^{-4\beta t} - e^{-2(\alpha+\beta)t} \end{aligned} \quad (11)$$

These probabilities are also shown in Figure 3c. The reason for the condition that "e_1 has not occurred" is because e_1 event can erase e_2 event (as we have discussed before).

Now the probability of the starting nucleotide A changing to G during time t, designated as p(G|A,t), is the summation of two probabilities. The first is 1/2 of the probability of $p(e_2 > 0, e_1 = 0,t)$ in Eq. (11) because the other 1/2 is for A to itself. The second is 1/4 of $p(e_1 > 0,t)$ in Eq. (10) because A→A, A→G, A→C and A→T each get ¼, so only 1/4 of $p(e_1 > 0,t)$ is for A→G. The summation of these two probabilities (Figure 3d) is p(G|A,t). This probability is equal to p(A|G,t), p(C|T,t), and p(T|C,t) in the K80 model. In other words, the summation of these two probabilities is the probability of a transition (P_s) during time t. Thus,

$$\begin{aligned} P_s &= \frac{p(e_2 > 0, e_1 = 0, t)}{2} + \frac{p(e_1 > 0, t)}{4} \\ &= \frac{e^{-4\beta t} - e^{-2(\alpha+\beta)t}}{2} + \frac{1 - e^{-4\beta t}}{4} \\ &= \frac{1}{4} + \frac{e^{-4\beta t}}{4} - \frac{e^{-2(\alpha+\beta)t}}{2} = P(G|A,t) = P(A|G,t) = P(T|C,t) = P(C|T,t) \end{aligned} \quad (12)$$

Similarly, the probability of the starting A changing to C (or to T) is 1/4 of $p(e_1 > 0,t)$ in Eq. (10) because 1/4 is for A→A, ¼ is for A→G and 1/4 is for A→T, so only 1/4 is for A→C (Figure 3d). This probability is the probability for a transversional change during time t,

$$P_v = \frac{p(e_1 > 0, t)}{4} = \frac{1 - e^{-4\beta t}}{4} \quad (13)$$

As a quick check of the derived transition probabil-

ities, we note that P_s and P_v are zero when t = 0 (or when = 0 and β = 0). This also implies that all diagonal elements in the transition probability matrix are equal to 1, and is what we have expected. When t = ∞, with α > 0 and β > 0, all entries in matrix P approaches ¼ (the equilibrium frequency of the K80 model). This is also what we expected.

For two aligned homologous sequences, P_s can be approximated by the proportion of sites differing by a transition (P), and $2P_v$ by the portion of sites differing by a transversion (Q, Figure 3e). Note that the expected Q is equal to $2P_v$ because there are two ways of having a transversional change. Therefore,

$$P = \frac{1}{4} + \frac{e^{-4\beta t}}{4} - \frac{e^{-2(\alpha+\beta)t}}{2} \quad (14)$$

$$Q = 2P_v = 2\left(\frac{1-e^{-4\beta t}}{4}\right) = \frac{1-e^{-4\beta t}}{2} \quad (15)$$

We can now first solve for βt from Eq. (15), and then substitute the solution for βt into Eq. (14) to solve for αt. This leads to

$$\begin{aligned} \alpha t &= -\frac{\ln(1-2P-Q)}{2} + \frac{\ln(1-2Q)}{4} \\ \beta t &= -\frac{\ln(1-2Q)}{4} \end{aligned} \quad (16)$$

Recall that evolutionary distance is defined as μt, where μ is the substitution rate which is equal to (α+2β) in the K80 model. Thus, the evolutionary distance based on the K80 model (D_{K80}) is (α+2β)t, which comes to

$$D_{K80} = \alpha t + 2\beta t = -\frac{\ln(1-2P-Q)}{2} - \frac{\ln(1-2Q)}{4} \quad (17)$$

Where P and Q can be approximated by the observed proportion of sites differing by a transition or a transversion from two aligned sequences, designated as P_{obs} and Q_{obs}. Similar to what I have mentioned with reference to D_{JC69}, P and Q may differ from P_{obs} and Q_{obs} even if the K80 model is followed during the sequence evolution. This is because 1) Limited aligned length of the two sequences may result in stochastic variation in P_{obs} and Q_{obs}, and 2) The two observed proportions may be distorted by alignment errors (i.e., misidentification of site homology). For example, two homologous sequences that have diverged for an infinite length of time according to the K80 model should have expected P and Q equal to 0.25 and 0.5, respectively. However, we may actually have P_{obs} > 0.25 or Q_{obs} > 0.5, which would render D_{K80} inapplicable. On the other hand, after sequence alignment and deletion of indels (because evolutionary distances are typically calculated without using sites with indels), P_{obs} and Q_{obs} may well be much smaller than the expected 0.25 and 0.5 leading to severer underestimation of the true distance. It is also possible to have P_{obs} and Q_{obs} values that, when used to replace P and Q in Eq. (16), result in negative αt or βt values that make no biological sense. The same applies to D_{JC69} (in fact to any evolutionary distances based on a substitution model). Methods for handling such situations are discussed later in the section on the GTR model.

We may optionally show D_{K80} in Eq. (17) to be a maximum likelihood estimator of the distance based on the K80 model, just like the K_{JC69} distance in Eq. (5). To see this, it is better to re-parameterize the K80 model by replacing αt and βt by D_{K80} and κ using the following relationship:

$$\begin{aligned} D_{K80} &= \alpha t + 2\beta t \\ \kappa &= \alpha t / \beta t \end{aligned} \quad (18)$$

Solving these two equations gives us

$$\begin{aligned} \alpha t &= \frac{D_{K80}\kappa}{\kappa+2} \\ \beta t &= \frac{D_{K80}}{\kappa+2} \end{aligned} \quad (19)$$

Substituting αt and βt into Eqs. (14) and (15) so that P and Q will be functions of D_{K80} and κ, and the likelihood function for deriving D_{K80} and κ is

$$\begin{aligned} L &= \left(\frac{1}{4}\right)^N P^{N_s} Q^{N_v} (1-P-Q)^{N-N_s-N_v} \\ \ln L &= N\ln\left(\frac{1}{4}\right) + N_s \ln P + N_v \ln Q + (N-N_s-N_v)\ln(1-P-Q) \end{aligned} \quad (20)$$

Where the constant term $N^*\ln(1/4)$ can be dropped in maximizing lnL to obtain the distance estimate, but need to be kept when performing likelihood ratio test for comparing different substitution models (e.g., K80 against TN93).

Taking partial derivatives with respect to D_{K80} and κ, setting them to zero and solving the simultaneous equations, we have

$$D_{K80} = -\frac{\ln(1-2P_{obs}-Q_{obs})}{2} - \frac{\ln(1-2Q_{obs})}{4} \quad (21)$$

$$\kappa = \frac{2\ln(1-2P_{obs}-Q_{obs})}{\ln(1-2Q_{obs})} - 1 \quad (22)$$

Where $P_{obs} = N_s/N$, and $Q_{obs} = N_v/N$. Using the two aligned sequences in Figure 2, we have N = 24 and P_{obs} = 4/24 and Q_{obs} = 2/24. These lead to D_{K80} = 0.3151, and κ = 4.9126. It may be relevant to add that, while D_{JC69} and D_{K80} are maximum likelihood estimates, distance formulae for F84 and TN93 models, obtained in the same way by equating the observed substitutions to expected substitutions, are generally not maximum likelihood estimates. This will become clear when we deal with these models.

We have previously derived the variance of D_{JC69} as the negative reciprocal of the second derivative of lnL with respect to D_{JC69}. This can be used only when the log-likelihood function is used to estimate a single parameter. When there are multiple parameters (e.g., D_{K80} and κ), we cannot use the same approach unless the parameters are not correlated. There are two common-

ly used methods for deriving variances of parameters. The first is the delta method [7-9], and the second uses the Fisher information matrix to obtain the variances and covariance matrix for the parameters. The delta method, which often yields nice and clean mathematical expressions for the variance, is illustrated in the Appendix. The method using the Fisher information matrix is shown below.

To estimate variance involving multiple parameters such as D_{K80} and κ, we first take the second order partial derivatives of lnL with respective to D_{K80} and κ, substituting the estimated D_{K80} and κ in Eqs. (21) and (22) into the second-order partial derivatives, arranging them into what is called a Fisher information matrix (M_{FI}) below, and compute the matrix inverse of M_{FI} (designated by M_{FI}^{-1}):

$$M_{FI} = \begin{bmatrix} -\frac{\partial^2 \ln L}{\partial \kappa^2} & -\frac{\partial^2 \ln L}{\partial \kappa \, \partial D_{K80}} \\ -\frac{\partial^2 \ln L}{\partial D_{K80} \partial \kappa} & -\frac{\partial^2 \ln L}{\partial D_{K80}^2} \end{bmatrix} \quad (23)$$

The diagonal elements of M_{FI}^{-1} are the variances for κ and D_{K80}, and the off-diagonal elements of M_{FI}^{-1} are covariances. The mathematical expression for the variance of κ is tedious, but that for the variance of D_{K80} is simpler:

$$V(D_{K80}) = \frac{a^2P + c^2Q - (aP + cQ)^2}{N}, \text{ where}$$
$$a = \frac{1}{1-2P-Q}, b = \frac{1}{1-2Q}, c = \frac{a+b}{2} \quad (24)$$

With the aligned sequences in Figure 2, we have N = 24 and empirical P = 4/24 and Q = 2/24. These lead to D_{K80} = 0.3151, and κ = 4.9126. The M_{FI} and M_{FI}^{-1} are

$$M_{FI} = \begin{bmatrix} 0.047435 & -0.286795 \\ -0.286795 & 49.641105 \end{bmatrix}$$
$$M_{FI}^{-1} = \begin{bmatrix} 21.84451668 & 0.126204037 \\ 0.126204037 & 0.020873724 \end{bmatrix} \quad (25)$$

Where the two parameters are in the order of κ and D_{K80}, i.e., the variance is 21.8445 for κ and 0.0209 for D_{K80}. The off-diagonal elements are covariances between the two parameters.

F84 and HKY85 model

The F84 and HKY85 model accommodate not only the differential substitution rates between transitions and transversions, but also different equilibrium nucleotide frequencies, in contrast to JC69 and K80 which assume equal equilibrium nucleotide frequencies. The same probabilistic reasoning used before can be applied to derive transition probabilities for the HKY80 model.

The rate matrix for the F84 model is

$$Q_{F84} = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{bmatrix} - & \beta\pi_G + \gamma\pi_G/\pi_R & \beta\pi_C & \beta\pi_T \\ \beta\pi_A + \gamma\pi_A/\pi_R & - & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & - & \beta\pi_T + \gamma\pi_T/\pi_Y \\ \beta\pi_A & \beta\pi_G & \beta\pi_C + \gamma\pi_C/\pi_Y & - \end{bmatrix} \end{array} \quad (26)$$

Where π_A, π_G, π_C and π_T are equilibrium frequencies, π_R and π_Y are frequencies of purines and pyrimidines, and the diagonal elements are constrained by each row summing up to 0. The parameter γ in Eq. (26) is sometimes replaced by κβ, but it is easier to understand the F84 model by using Q_{F84} specified in Eq. (26).

We may view the F84 model as featuring two events (e_1 and e_2). Suppose we start with a nucleotide A. Event e_1 occurs with rate β. When it occurs, the original A will be replaced by a nucleotide drawn randomly from a nucleotide pool in which the nucleotide frequencies are the same as the equilibrium frequencies. This means that the original A has a rate of $\beta\pi_A$, $\beta\pi_G$, $\beta\pi_C$ and $\beta\pi_T$ to change to A, G, C and T, respectively, when e_1 occurs. This is different from the K80 model where, when e_1 occurs, the original A has a rate of 0.25 to change to any of the four nucleotides. Event e_2 has a rate of γ to occur, and will result in the original A being replaced by a purine drawn randomly from a purine pool with A and G frequencies specified as $\pi_A/_R$ and $\pi_G/_R$. Thus, the original A has a rate of $\gamma\pi_A/\pi_R$ and $\gamma\pi_G/\pi_R$ to change to A and G when e_2 occurs. This again differs from K80 where the original A has a rate of 0.5 of changing to A or Gwhen e_2 occurs. These events are illustrated in Figure 4a, where we use x to represent $\beta+\gamma/\pi_R$. Note that it is not a good idea to use α to represent $\beta+\gamma/\pi_R$ for two reasons. First, if we had started with a nucleotide C or T instead of A, then we would have $\beta+\gamma/\pi_Y$ instead of $\beta+\gamma/\pi_R$ which would force us to use α_1 and α_2 to distinguish between the two. A casual reader will then be misled to think that F84 has three rate parameters (i.e., β, α_1 and α_2) without knowing that α_1 and α_2 are used as different functions of the same rate parameter γ. Second, I have reserved α to represent β+γ in Figure 4b which simplifies the derivation of transition probabilities illustrated in Figure 4.

Note that whenever event e_1 happens, the original A is replace by A, C, G and T with probabilities π_A, π_G, π_C and π_T, no matter how many e_2 events has occur before or after the occurrence of e_1. This is similar to the scenario involving the K80 model, except that the K80 model assumes equal nucleotide frequencies. It might help to think of a long sequence with L sites being A at time 0. If these L sites each have experienced at least one e_2 event, then these sites will either be A or G with probabilities π_A and π_G, respectively, and we expect to have π_AL sites being A and π_GL sites being G. In contrast, if each of these L sites has experienced at least one e_1 event, then the site will be replaced by A, C, G, or T with probabilities π_A, π_G, π_C and π_T, and we expect to observe A, C, G and T in π_AL, π_GL, π_CL, and π_TL sites, respectively. Any number of e_2 events occurring before or after the e_1 event does not change this expectation. This means that e_1 erases e_2, but not vice versa. The occurrence of an e_2 event is informative only when no e_1 event has happened.

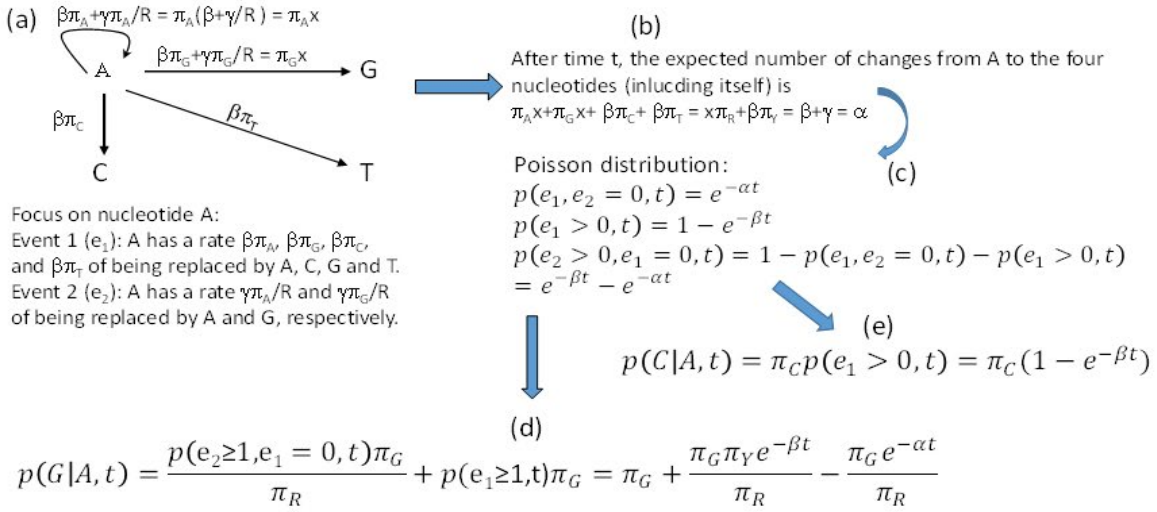

Figure 4: Derivation of transition probabilities based on the F84 model. π_A, π_G,π_C, π_T and π_R, π_Y are equilibrium frequencies of A, C, G, T, purine (A + G) and pyrimidine (C + T). Event e_1 occurs at a rate of β and leads to the original A being replaced by any of the four nucleotides according to their equilibrium frequencies, and event e_2 occurs at a rate of γ and results in the original A being replaced by either A or G according to their frequencies in the purine pool, i.e., π_A/π_R and π_G/π_R. I used x as a shorthand for $\beta+\gamma/\pi_R$. The rate γ does not appear in the final transition probabilities because it has been absorbed into α which equals β+γ shown in (b). P(j|i,t) means the probability of changing from the original nucleotide i to nucleotide j after time t, and is synonymous to $p_{ij}(t)$ or simply p_{ij} in this paper.

After time t, the total flow of the original A to the four nucleotides (including itself, Figure 4a and Figure 4b) is

$$\pi_A x+\pi_G x+\beta\pi_C+\beta\pi_T = \pi_R x+\pi_Y\beta = \pi_R(\beta+\gamma/\pi_R)+\pi_Y\beta = \beta+\gamma = \alpha \quad (27)$$

So the probability that no substitution has happened during time t (Figure 4c), according to Poisson distribution, is

$$p(e_1, e_2 = 0, t) = e^{-\alpha t} \quad (28)$$

The rate of A changing to A, G, C, and T through e_1 is $\beta\pi_A + \beta\pi_G + \beta\pi_C + \beta\pi_T = \beta$, so the probability that at least one e_1 has occurred during time t is

$$p(e_1 > 0, t) = 1 - e^{-\beta t} \quad (29)$$

The probability that e_2 has happened but e_1 has not is then

$$p(e_2 > 0, e_1 = 0, t) = 1 - p(e_1, e_2 = 0, t) - p(e_1 > 0, t) = e^{-\beta t} - e^{-\alpha t} \quad (30)$$

The reason for the condition that "e_1 has not occurred" is because e_1 event can erase e_2 event. With these, it is easy to derive transition probability from A to G (Figure 4d) as the summation of 1) A fraction of π_G of p(e1 > 0,t), which is the probability of e_1 event that results in the original A being replaced by A, C, G, and T with probabilities π_A, π_G, π_C and π_T, and 2) A fraction of π_G/π_R of $p(e_2 > 0, e_1 = 0, t)$, which is the probability that e_2 events not erased by e_1. That is,

$$p(G|A,t) = p(e_1 > 0, t)\pi_G + \frac{p(e_2 > 0, e_1 = 0, t)\pi_G}{\pi_R} = \pi_G + \frac{\pi_G\pi_Y e^{-\beta t}}{\pi_R} - \frac{\pi_G e^{-\alpha t}}{\pi_R} \quad (31)$$

From now on, p(j|i,t) will be written simply as p_{ij}, so p(G|A,t) is p_{AG}. With the same reasoning, we can derive transition probabilities for other A↔G and C↔T substitutions. Note that the two rate parameters in the F84 model (β and γ) have been re-parameterized into α (= β + γ) and β in Eq. (31). The transition probability from the original A to C (a transversion, Figure 4e) is simply

$$p_{AC} = \pi_C p(e_1 > 0, t) = \pi_C(1 - e^{-\beta t}) \quad (32)$$

For other transversions, e.g., p_{AT}, one just need to replace π_C by π_T. The complete transition probability matrix for the F84 model is

$$P_{F84} = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{bmatrix} \pi_A+\pi_A\pi_Y x_1+\pi_G x_2 & \pi_G+\pi_G\pi_Y x_1-\pi_G x_2 & \pi_C(1-e^{-\beta t}) & \pi_T(1-e^{-\beta t}) \\ \pi_A+\pi_A\pi_Y x_1-\pi_A x_2 & \pi_G+\pi_G\pi_Y x_1+\pi_A x_2 & \pi_C(1-e^{-\beta t}) & \pi_T(1-e^{-\beta t}) \\ \pi_A(1-e^{-\beta t}) & \pi_G(1-e^{-\beta t}) & \pi_C+\pi_C\pi_R x_3+\pi_G x_4 & \pi_T+\pi_T\pi_R x_3-\pi_T x_4 \\ \pi_A(1-e^{-\beta t}) & \pi_G(1-e^{-\beta t}) & \pi_C+\pi_C\pi_R x_3-\pi_C x_4 & \pi_T+\pi_T\pi_R x_3+\pi_C x_4 \end{bmatrix} \quad (33)$$

Where

$$x_1 = \frac{e^{-\beta t}}{\pi_R}, x_2 = \frac{e^{-\alpha t}}{\pi_R}, x_3 = \frac{e^{-\beta t}}{\pi_Y}, x_4 = \frac{e^{-\alpha t}}{\pi_Y} \quad (34)$$

As a quick check of the transition probabilities, we first note that when t = 0 (or when α = 0 and β = 0), then the diagonal elements are 1 and all off-diagonal elements are 0, which is what we expected. Second, when t = ∞ with α > 0 and β > 0, then the transition probabilities will approach the equilibrium frequencies, which is also what we expected.

To obtain the distance for the F84 model (D_{F84}), recall that a distance is defined as μt where μ is the average substitution rate, i.e., substitution rates in Eq. (26) weighted by the equilibrium frequencies:

$$D_{F84} = 2\pi_A\pi_G(\beta t+\gamma t/\pi_R)+2\pi_T\pi_C(\beta t+\gamma t/\pi_Y)+2\pi_Y\pi_R\beta t \quad (35)$$

Now we need to obtain βt and γt in order to calculate D_{F84}. We can obtain αt and βt, and then obtain γt = αt - βt, remembering that α = β + γ (Figure 4b and Eq. (27)]. The method we will use is the same as that for the K80 model, i.e., we obtain the expected transitions and transversions, designated E(S) and E(V), respectively, from transition probabilities and equate them to the

observed S and V to solve for αt and βt. With the property of time reversibility (e.g., $\pi_A \cdot p_{AG} = \pi_G \cdot p_{GA}$), we have

$$\begin{aligned} E(S) &= 2\pi_A p_{AG} + 2\pi_C p_{CT} \\ E(V) &= 2\pi_A p_{AT} + 2\pi_A p_{AC} + 2\pi_G p_{GC} + 2\pi_G p_{GT} \end{aligned} \quad (36)$$

Equating E(S) and E(V) to the observed S and V, and solving these two equations with the two unknowns (αt and βt), we have

$$\alpha t = \ln\left(\frac{-2(\pi_A\pi_G\pi_R\pi_Y^2 + \pi_C\pi_T\pi_R^2\pi_Y)}{S\pi_R^2\pi_Y^2 - 2\pi_A\pi_G\pi_R\pi_Y^2 - 2\pi_C\pi_T\pi_R^2\pi_Y + (\pi_A\pi_G\pi_Y^2 + \pi_C\pi_T\pi_R^2)V}\right) \quad (37)$$

$$\beta t = -\ln\left(1 - \frac{V}{2\pi_R\pi_Y}\right) \quad (38)$$

Substitute βt and γt (= αt - βt) into Eq. (35) and, after some algebraic manipulation, we have a more useful form of D_{F84}:

$$D_{F84} = \frac{2}{\pi_R\pi_Y}\left[-(\pi_A\pi_G + \pi_C\pi_T)\pi_R\pi_Y \ln(x_1) + \pi_C\pi_T\pi_R \ln\left(\frac{x_2}{x_3}\right) + \pi_A\pi_G\pi_Y \ln\left(\frac{x_2}{x_3}\right) - \pi_R^2\pi_Y^2 \ln(x_1)\right] \quad (39)$$

Where

$$\begin{aligned} x_1 &= 1 - \frac{V}{2\pi_R\pi_Y}; \\ x_2 &= (\pi_A\pi_G\pi_Y + \pi_C\pi_T\pi_R)(2\pi_R\pi_Y - V) \\ x_3 &= -S\pi_R^2\pi_Y^2 + 2\pi_A\pi_G\pi_R\pi_Y^2 + 2\pi_C\pi_T\pi_Y\pi_R^2 - \pi_A\pi_G\pi_Y^2 V - \pi_C\pi_T\pi_R^2 V \end{aligned} \quad (40)$$

To illustrate the calculation of D_{F84}, we may use the two aligned sequences in Figure 2 which gives us π_A = 6/48, π_C = 12/48, π_G = 10/48, π_T = 20/48, S = 4/24, V = 2/24, αt = 0.5778363341, βt = 0.2076393648, γt = αt - βt = 0.3701969693, and D_{F84} = 0.3198867427. The variance of the D_{F84} can be obtained by either the delta method or the method using Fisher information matrix.

A substitution model similar to the F84 model is the HKY85 model, with its rate matrix specified as:

$$Q_{HKY85} = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{bmatrix} - & (\beta+\gamma)\pi_G & \beta\pi_C & \beta\pi_T \\ (\beta+\gamma)\pi_A & - & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & - & (\beta+\gamma)\pi_T \\ \beta\pi_A & \beta\pi_G & (\beta+\gamma)\pi_C & - \end{bmatrix} \end{array} \quad (41)$$

Where (β+γ) is often written as α and the diagonal elements are constrained by each row summing up to 0. The HKY85 model and the F84 model differ only in the specification of rates involving transitions. q_{AG} and q_{CT} are $\pi_G(\beta+\gamma)$ and $\pi_T(\beta+\gamma)$ in the HKY85 model specified in Eq. (41), in contrast to $\pi_G(\beta+\gamma/\pi_R)$ and $\pi_T(\beta+\gamma/\pi_Y)$, respectively, in the F84 model specified in Eq. (26). By comparing these rates, it becomes obvious that the F84 model would be equivalent to the HKY85 model if $\pi_R = \pi_Y$.

We can obtain the transition probabilities for the HKY85 model in the same way as that for the F84 model. In short, we again start with a nucleotide A and envision two events e_1 and e_2. Event e_1 occurs with rate β, and results in the original A replaced by any of the four nucleotides with probabilities equal to their respective equilibrium frequencies. Event e_2 occurs with a rate γ and results in the original A being replaced by either A or G with the probabilities equal to their respective equilibrium frequencies. Fictionalized in this way, the expected number of substitutions after time t is $\beta(\pi_A+\pi_G+\pi_C+\pi_T) + \gamma(\pi_A+\pi_G) = \beta+\gamma R$. According to the Poisson distribution, the probability that no substitution has happened during time t is

$$p(e_1, e_2 = 0, t) = 1 - e^{-(\beta+\gamma R)} \quad (42)$$

The probability that at least one e_1 occurred after time t is

$$p(e_1 > 0, t) = 1 - e^{-\beta t} \quad (43)$$

The probability that e_2 has occurred but e_1 has not is

$$p(e_2 > 0, e_1 = 0, t) = 1 - p(e_1, e_2 = 0, t) - p(e_1 > 0, t) = e^{-\beta t} - e^{-(\beta+\gamma R)} \quad (44)$$

The transition probability p(G|A,t), abbreviated as p_{AG}, is

$$p_{AG} = \pi_G p(e_1 > 0, t) + \frac{\pi_G}{\pi_R} p(e_2 > 0, e_1 = 0, t) = \pi_G + \frac{\pi_G\pi_Y e^{-\beta t}}{\pi_R} - \frac{\pi_G e^{-(\beta+\pi_R\gamma)t}}{\pi_R} \quad (45)$$

In the same way, we can derive other transition probabilities which are shown below:

$$P_{HKY} = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{bmatrix} \pi_A + \pi_A x_1 + \pi_G x_2 & \pi_G + \pi_G x_1 - \pi_G x_2 & \pi_C(1-e^{-\beta t}) & \pi_T(1-e^{-\beta t}) \\ \pi_A + \pi_A x_1 - \pi_A x_2 & \pi_G + \pi_G x_1 + \pi_A x_2 & \pi_C(1-e^{-\beta t}) & \pi_T(1-e^{-\beta t}) \\ \pi_A(1-e^{-\beta t}) & \pi_G(1-e^{-\beta t}) & \pi_C + \pi_C x_3 + \pi_T x_4 & \pi_T + \pi_T x_3 - \pi_T x_4 \\ \pi_A(1-e^{-\beta t}) & \pi_G(1-e^{-\beta t}) & \pi_C + \pi_C x_3 - \pi_C x_4 & \pi_T + \pi_T x_3 + \pi_C x_4 \end{bmatrix} \end{array} \quad (46)$$

Where

$$x_1 = \frac{\pi_Y e^{-\beta t}}{\pi_R}; x_2 = \frac{e^{-(\beta+\pi_R\gamma)t}}{\pi_R}; x_3 = \frac{\pi_R e^{-\beta t}}{\pi_Y}; x_4 = \frac{e^{-(\beta+\pi_Y\gamma)t}}{\pi_Y} \quad (47)$$

As a quick check of the transition probabilities, we first note that when t = 0 (or when α = 0 and β = 0), then the diagonal elements are 1 and all off-diagonal elements are 0, which is what we expected. Second, when t approaches infinity with β > 0 and γ > 0, then the transition probabilities will approach the equilibrium frequencies, which is also what we expected.

We cannot derive the distance for the HKY85 model by following the same approach as that for the F84 model. Hasegawa, et al. [4] has tried this approach but were not successful because there is no explicit solution for βt and γt. However, if we treat the A↔G transition and C↔T transition separate, then we can solve for βt and γt [10]. In other words, we obtain one set of βt and γt from observed A↔G transitions and transversions, and another set of βt and γt from observed C↔T transitions and transversions. βt in the two sets are the same as that in Eq. (38), but γt is different between the two sets of estimates. We can then take a weighted average of γt. Admittedly, this does sound mathematically clumsy and explains why HKY85, while commonly used in phylogenetic analysis involving a likelihood framework or Bayesian inference, is almost never used in distance-based phylogenetics.

Here is the somewhat circuitous protocol to get βt and γt from HKY85. The expected numbers of A↔G

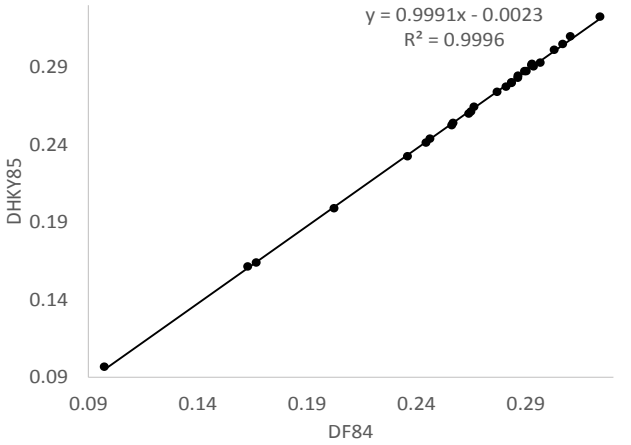

Figure 5: Evolutionary distances from the HKY85 and F84 models are nearly identical.

and C↔T transitions, designated S_R and S_Y, respectively, and transversions are

$$\begin{aligned} E(S_R) &= 2\pi_A p_{AG} \\ E(S_Y) &= 2\pi_C p_{CT} \\ E(V) &= 2\pi_A p_{AT} + 2\pi_A p_{AC} + 2\pi_G p_{GC} + 2\pi_G p_{GT} \end{aligned} \quad (48)$$

Setting $E(S_R)$ and E(V) to their the observed S_R and V, and solve for βt and γt, we have

$$\begin{aligned} \beta t &= -\ln\left(1-\frac{V}{2\pi_R\pi_Y}\right) \\ \gamma_R t &= \frac{1}{\pi_R}\ln\left(\frac{\pi_A\pi_G(2\pi_R\pi_Y - V)}{2\pi_A\pi_G\pi_R\pi_Y - S_R\pi_Y\pi_R^2 - \pi_A\pi_G\pi_Y V}\right) \end{aligned} \quad (49)$$

Where βt is the same as that in Eq. (38), and $\gamma_R t$ in Eq. (49) is γt estimated from observed S_R and V.

Now we obtain another set of solutions for βt and γt by setting $E(S_Y)$ and E(V) to their observed S_Y and V, and solve for βt and γt, we have the same βt but a different γt:

$$\gamma_Y t = \frac{1}{\pi_Y}\ln\left(\frac{\pi_C\pi_T(2\pi_R\pi_Y - V)}{2\pi_C\pi_T\pi_R\pi_Y - S_Y\pi_R\pi_Y^2 - \pi_C\pi_T\pi_R V}\right) \quad (50)$$

A weighted average of γt could be

$$\gamma t = \pi_R\gamma_R t + \pi_Y\gamma_Y t \quad (51)$$

The distance for the HKY model

$$D_{HKY85} = \mu t = 2\pi_A\pi_G(\beta t+\gamma t) + 2\pi_T\pi_C(\beta t+\gamma t) + 2\pi_Y\pi_R\beta t \quad (52)$$

To compute D_{HY85} using the two aligned sequences in Figure 2, we have π_A = 6/48, π_C = 12/48, π_G = 10/48, π_T = 20/48, S_Y = 4/24, S_R = 0, V = 2/24, βt = 0.2076393648, $\gamma_R t$ = -0.2223239164, $\gamma_Y t$ = 1.047432870, weighted γt = 0.624180608, D_{HKY85} = 0.308904. I intentionally choose the aligned sequences in Figure 2 with S_R = 0 just to see if D_{HKY85} would behave strangely. It did not. For comparison, the same two sequences yield D_{F84} = 0.319887.

In general, D_{HKY85} is slightly smaller than D_{F84}. I used the eight vertebrate COI sequences in the FASTA file VertCOI.fas that comes with DAMBE [11] to compute both D_{HKY85} and D_{F84} (Figure 5). The difference is minor, although D_{HKY85} is consistently but slightly smaller than D_{F84}.

The HKY85 model itself may not carry much biological significance given the existence of the F84 model. However, the twists involved in computing the evolutionary distance, i.e., the separate estimation of $\gamma_{A\leftrightarrow G}$ and $\gamma_{C\leftrightarrow T}$, lead very naturally to a very useful TN93 model that we will cover next.

TN93 model

We have come far, so far that we need hardly any extra effort to derive transition probabilities for the TN93 model. There are two equivalent specifications of the rate matrix for the TN93 model. The first is

$$Q_{TN93} = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array}\begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{bmatrix} - & \beta\pi_G + \gamma_R\pi_G/\pi_R & \beta\pi_C & \beta\pi_T \\ \beta\pi_A + \gamma_R\pi_A/\pi_R & - & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & - & \beta\pi_T + \gamma_Y\pi_T/\pi_Y \\ \beta\pi_A & \beta\pi_G & \beta\pi_C + \gamma_Y\pi_C/\pi_Y & - \end{bmatrix} \end{array} \quad (53)$$

Where the diagonal elements are constrained by each row summing up to 0. The second specification simply replaces $(\beta + \gamma_R/\pi_R)$ by α_1 and $(\beta + \gamma_Y/\pi_Y)$ by α_2. We see that TN93 is reduced to F84 if $\gamma_R = \gamma_Y$, and to HKY85 if $\gamma_R / \pi_R = \gamma_Y / \pi_Y$.

The similarity between TN93 and F84 allows us to re-use Figure 4 for deriving transition probabilities for TN93. We only need to add a subscript R to γ and α in Figure 4 so that we have γ_R and α_R as rates for purine, keeping everything else the same, and we instantly obtain the transition probabilities for transitional substitutions between purines and for transversional substitutions as shown in Figure 4. To get transition probabilities between pyrimidines, we can just replace the original nucleotide A in Figure 4 by nucleotide C or T and rename γ and α in Figure 4 to γ_Y and α_Y. Note that our $\alpha_R = \beta+\gamma_R$, and $\alpha_Y = \beta+\gamma_Y$.

The transition probability matrix for the TN93 model is

$$P_{TN93} = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array}\begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{bmatrix} \pi_A + \pi_A\pi_Y x_1 + \pi_G x_2 & \pi_G + \pi_G\pi_Y x_1 - \pi_G x_2 & \pi_C(1-e^{-\beta t}) & \pi_T(1-e^{-\beta t}) \\ \pi_A + \pi_A\pi_Y x_1 - \pi_A x_2 & \pi_G + \pi_G\pi_Y x_1 + \pi_A x_2 & \pi_C(1-e^{-\beta t}) & \pi_T(1-e^{-\beta t}) \\ \pi_A(1-e^{-\beta t}) & \pi_G(1-e^{-\beta t}) & \pi_C + \pi_C\pi_R x_3 + \pi_G x_4 & \pi_T + \pi_T\pi_R x_3 - \pi_T x_4 \\ \pi_A(1-e^{-\beta t}) & \pi_G(1-e^{-\beta t}) & \pi_C + \pi_C\pi_R x_3 - \pi_C x_4 & \pi_T + \pi_T\pi_R x_3 + \pi_C x_4 \end{bmatrix} \end{array} \quad (54)$$

Where x_1 and x_3 are the same as those in Eq. (34), but x_2 has α replaced by α_R and x_4 has α replaced by α_Y, i.e.,

$$x_1 = \frac{e^{-\beta t}}{\pi_R}, x_2 = \frac{e^{-\alpha_R t}}{\pi_R}, x_3 = \frac{e^{-\beta t}}{\pi_Y}, x_4 = \frac{e^{-\alpha_Y t}}{\pi_Y} \quad (55)$$

To obtain the distance for the TN93 model (D_{TN93}), recall that a distance is defined as μt where μ is the average substitution rate, i.e., substitution rates in Eq. (53) weighted by the equilibrium frequencies, so:

$$D_{TN93} = 2\pi_A\pi_G(\beta t+\gamma_R t/\pi_R) + 2\pi_T\pi_C(\beta t+\gamma_Y t/\pi_Y) + 2\pi_Y\pi_R\beta t \quad (56)$$

Now we need to obtain $\alpha_R t$, $\alpha_Y t$, and βt. The method we will use is the same as that for the K80 and F84 models, i.e., we obtain the expected numbers of A↔G transitions, C↔T transitions, and transversions, designated $E(S_R)$, $E(S_Y)$ and E(V), respectively, from transition probabilities, and equate them to the observed S_R, S_Y

and V to solve for $\alpha_R t$, $\alpha_Y t$, and βt:

$$\begin{aligned} E(S_R) &= 2\pi_A p_{AG} = S_R \\ E(S_Y) &= 2\pi_C p_{CT} = S_Y \\ E(V) &= 2\pi_A p_{AT} + 2\pi_A p_{AC} + 2\pi_G p_{GC} + 2\pi_G p_{GT} = V \end{aligned} \quad (57)$$

The resulting $\alpha_R t$, $\alpha_Y t$, and βt are

$$a_R t = \ln\left(\frac{2\pi_A\pi_G\pi_R\pi_Y}{2\pi_A\pi_G\pi_R\pi_Y - \pi_R^2\pi_Y S_R - \pi_A\pi_G\pi_Y V}\right) \quad (58)$$

$$a_Y t = \ln\left(\frac{2\pi_C\pi_T\pi_R\pi_Y}{2\pi_C\pi_T\pi_R\pi_Y - \pi_Y^2\pi_R S_Y - \pi_C\pi_T\pi_R V}\right) \quad (59)$$

$$\beta t = -\ln\left(1 - \frac{V}{2\pi_R\pi_Y}\right) \quad (60)$$

If one wishes to express D_{TN93} in S_R, S_Y and V, then one may just substitute $\gamma_R t$, $\gamma_Y t$, and βt into Eq. (56), which yields:

$$D_{TN93} = \frac{2\pi_A\pi_G[\pi_Y \ln(x_1) + \ln(x_2)]}{\pi_R} + \frac{2\pi_C\pi_T[\pi_R \ln(x_1) + \ln(x_3)]}{\pi_Y} - 2\pi_R\pi_Y x_1 \quad (61)$$

Where

$$\begin{aligned} x_1 &= 1 - \frac{V}{2\pi_R\pi_Y}; \\ x_2 &= \frac{2\pi_A\pi_G\pi_R\pi_Y}{2\pi_A\pi_G\pi_R\pi_Y - S_R\pi_R^2\pi_Y - \pi_A\pi_G\pi_Y V}; \\ x_3 &= \frac{2\pi_C\pi_T\pi_R\pi_Y}{2\pi_C\pi_T\pi_R\pi_Y - S_Y\pi_Y^2\pi_R - \pi_C\pi_T\pi_R V}. \end{aligned} \quad (62)$$

To illustrate the application of D_{TN93} with the two aligned sequences in Figure 2, we have $\pi_A = 6/48$, $\pi_C = 12/48$, $\pi_G = 10/48$, $\pi_T = 20/48$, $S_Y = 4/24$, $S_R = 0$, $V = 2/24$, $\alpha_R t = 0.13353$, $\alpha_Y t = 0.90593$, $\beta t = 0.20764$, $\gamma_R t = \alpha_R t - \beta t = -0.07411$, $\gamma_Y t = \alpha_R t - \beta t = 0.69829$, $D_{TN93} = 0.35299$. The variance of the D_{TN93} can be obtained by either the delta method or the method using Fisher information matrix. Note that $S_R = 0$ means no information for estimating $\alpha_R t$ properly.

I should mention that all distance formulations in this paper are known as Independently Estimated (IE) distances because they use information from only two aligned sequences and are independent of other pairs of sequences. Practical molecular phylogenetic analysis typically would use Simultaneously Estimated (SE) distances [12,13] which use information from all pairs of sequences. SE distances are implemented in MEGA [14] and DAMBE [11,15]. The PhyPA [16] function in DAMBE, which performs phylogenetic reconstruction base on pairwise alignment when reliable multiple sequence alignment is difficult to obtain for highly diverged sequences, uses SE distances only.

In short, the approach of deriving transition probabilities by probability reasoning can go a long way if one can do good bookkeeping. In particularly, the probability reasoning approach is very useful for conceptual understanding. However, the approach becomes increasing difficult with more complicated substitution models. Two alternative approaches, one involving solving differential equations and the other involving matrix exponential and logarithms, are often used in practical computation with the GTR model for nucleotide sequences and amino acid-based substitution models. They will be numerically illustrated elsewhere.

Acknowledgements

This study is funded by the Discovery Grant from Natural Science and Engineering Research Council of Canada (RGPIN/261252-2013). I thank C. Vlasschaert and S. Aris-Brosou for feedback.

References

1. Felsenstein J (2004) Inferring phylogenies. Sinauer, Sunderland, Massachusetts, 664.
2. Yang Z (2006) Computational molecular evolution. Oxford University Press, Oxford, 357.
3. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512-526.
4. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160-174.
5. Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN, Mammalian protein metabolism. Academic Press, New York, 21-132.
6. Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16: 111-120.
7. Kimura M, Ohta T (1972) On the stochastic model for estimation of mutational distance between homologous proteins. J Mol Evol 2: 87-90.
8. Waddell PJ, Steel MA (1997) General time-reversible distances with unequal rates across sites: mixing gamma and inverse Gaussian distributions with invariant sites. Mol Phylogenet Evol 8: 398-414.
9. Xia X (2007) Bioinformatics and the cell: Modern computational approaches in genomics, proteomics and transcriptomics. Springer US, New York, 349.
10. Rzhetsky A, Nei M (1995) Tests of applicability of several substitution models for DNA sequence data. Mol Biol Evol 12: 131-151.
11. Xia X (2013) DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol 30: 1720-1728.
12. Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci U S A 101: 11030-11035.
13. Xia X (2009) Information-theoretic indices and an approximate significance test for testing the molecular clock hypothesis with genetic distances. Mol Phylogenet Evol 52: 665-676.
14. Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33: 1870-1874.
15. Xia X (2017) DAMBE6: New tools for microbial genomics, phylogenetics and molecular evolution. J Hered 108: 431-437.
16. Xia X (2016) PhyPA: Phylogenetic method with pairwise sequence alignment outperforms likelihood methods in phylogenetics involving highly diverged sequences. Mol Phylogenet Evol 102: 331-343.