Evolutionary distances corrected for purifying selection and ancestral polymorphisms

Gonzalo Oteo–García

Department of Biological Sciences, School of Applied Sciences,
University of Huddersfield, Queensgate, Huddersfield, UK

José–Angel Oteo*

Departamento de Física Teórica, Universidad de Valencia,
and Institute for Integrative Systems Biology (I2SysBio),
46100-Burjassot, Valencia, Spain

Abstract

Evolutionary distance formulas that take into account effects due to ancestral polymorphisms and purifying selection are obtained on the basis of the full solution of Jukes–Cantor and Kimura DNA substitution models. In the case of purifying selection two different methods are developed. It is shown that avoiding the dimensional reduction implicitly carried out in the conventional model solving is instrumental to incorporate the quoted effects into the formalism. The problem of estimating the numerical values of the model parameters, as well as those of the correction terms, is not addressed.

Keywords: molecular clock, DNA substitution model, evolutionary distance, ancestral polymorphisms, purifying selection.
2010 MSC: 92D15
PACS: 87.23.-Kg

1. Introduction

Molecular evolutionary clock models are used to date evolutionary events [1, 2, 3, 4]. They provide a deterministic law that rules, under appropriate hypotheses, the time–evolution followed by the base proportions of molecular sequences subjected to genetic mutations. The question faced is actually a difficult inverse problem that poses as follows. Starting from two experimentally measured molecular sequences we want to ascertain the time elapsed since their divergence with the rationale that the base substitution rates have remained constant.

When a molecular clock model has all the parameters tuned, the mathematical solution tracks the instantaneous values of the nucleotide sequence base proportions, whenever the initial base proportions are given as input. As this information is experimentally unreachable we face an ill–posed problem in forward time direction whereas the nature of experimental data rather fits to a backward time analysis.

In this article we focus our attention on the original Kimura three substitution type model (K3ST) [5] and extend it to include effects due to purifying selection and ancestral polymorphisms. Firstly, we solve *ab initio* the differential equation system of the model in the matrix (Markov) formalism as presented in [6]. In this scheme one can solve at

*Corresponding author
Email address: oteo@uv.es (José–Angel Oteo)

once the history of the base proportions of a nucleotide sequence and the substitutions that appear when two homologous nucleotide sequences are compared. The solutions of the first problem may be represented as trajectories in a phase space that has dimension four (with three degrees of freedom). The second one has dimension sixteen (with fifteen degrees of freedom) and introducing a symmetry consideration it is reduced to dimension ten. The first problem is relevant for maximum likelihood and Bayesian computations whereas the second is essential to deal with evolutionary distance evaluation.

The four dimensional solution of the molecular clock originally obtained by Kimura establishes a relationship between the number of substitutions observed in two homologous nucleotide sequences and the time elapsed since their divergence. This solution has to be seen as a projection of the full ten–dimensional solution onto a subspace of dimension four. The projection takes place prior to solving the full system of differential equations, a fact that renders easier the algebraic resolution. This four dimensional solution must not be mistaken with the four dimensional solution that gives the instantaneous base proportions in a molecular sequence along the time, i.e. the first problem mentioned above.

Here we solve the mathematical problem in phase space of dimension ten. Of course, the four dimensional solution obtained by Kimura is easily recovered, as well as his celebrated evolutionary distance formula. The reason to choose working in the larger phase space is to allow the introduction of new elements in the substitution rates model. In particular we study the presence of ancestral polymorphisms [7, 8, 9] and the phenomenon of purifying selection [9, 10]. These are two different mechanisms proposed to explain the observed apparent acceleration of substitution rates in short times. We will show how the matrix formalism accommodates both situations in a natural way. In particular, we retrieve and generalize a recent result on ancestral polymorphisms [8]. Eventually, we introduce two approaches for removing the effect of deleterious mutations. The final output consists of a generalization of the K3ST evolutionary distance formula that takes into account the combined effect of both ancestral polymorphisms and purifying selection.

The corresponding generalized results for Kimura two parameter [11] and Jukes–Cantor (JC) [12] substitution models appear by appropriately equating parameters in K3ST model. As a matter of fact, we will particularize the generalized results only for JC model.

Our intention has been not to save technical details as they may be of help to extend these results to higher level molecular clock models.

2. Molecular clock matrix formalism

In the forthcoming five Sections, time reference $t = 0$ will correspond to the point when lineages start to diverge and thus modern time is $t > 0$. In Sections 7 to 9, $t = 0$ will refer to modern time.

We denote the four bases T, C, A, G, as $1, 2, 3, 4$, in the matrix formalism we use. Following [6] we introduce three different matrices, each one with specific purposes:

Matrix of Substitution Rates. This matrix defines the substitution model. Given a nucleotide sequence, h_{ij}, $i \neq j$, stands for the rate of substitution of base i by base j. The matrix elements in the diagonal are $h_{ii} = -\sum_{j \neq i} h_{ji} < 0$. Thus, the sum of every single column vanishes. In particular, K3ST model (see Figure 1) is defined by the symmetric transition matrix

$$H = \begin{pmatrix} -\Omega & \alpha & \beta & \gamma \\ \alpha & -\Omega & \gamma & \beta \\ \beta & \gamma & -\Omega & \alpha \\ \gamma & \beta & \alpha & -\Omega \end{pmatrix} \tag{1}$$

where $\Omega = \alpha + \beta + \gamma$. JC model deals with just one rate, μ, and it is obtained from Kimura's with the replacements $\alpha = \beta = \gamma \equiv \mu$. Notice that α, β, γ, and μ, take non–negative values because they represent substitution rates.

Divergence Matrix: $X(t) = (x_{ij}(t))$, $i, j = 1, \ldots, 4$. This matrix involves nucleotide sequences of two lineages and solves the time–evolution problem in the sense that it provides a link between the theoretical framework and the experimentally measurable data. Given a site, $x_{ij}(t)$ stands for the probability at time t that in the first nucleotide sequence the base is i and in the second one the base is j. The probability conservation conveys $\sum_{i,j} x_{i,j}(t) = 1$, for all time t. These matrix elements correspond to the sixteen substitution probability functions in Table 1, namely

$$X(t) = \begin{pmatrix} S_1 & P_1 & Q_1 & R_1 \\ \bar{P}_1 & S_2 & R_2 & Q_2 \\ \bar{Q}_1 & \bar{R}_2 & S_3 & P_2 \\ \bar{R}_1 & \bar{Q}_2 & \bar{P}_2 & S_4 \end{pmatrix}. \tag{2}$$

The comparison of two experimental homologous nucleotide sequences allows to assign numerical estimates to the matrix elements of X.

Conventional wisdom establishes that at time $t = 0$ (*i.e.*, *coalescence*)

$$X(t = 0) = \begin{pmatrix} \pi_T(0) & 0 & 0 & 0 \\ 0 & \pi_C(0) & 0 & 0 \\ 0 & 0 & \pi_A(0) & 0 \\ 0 & 0 & 0 & \pi_G(0) \end{pmatrix}, \tag{3}$$

where $\pi_T(0) + \pi_C(0) + \pi_A(0) + \pi_G(0) = 1$, and $\pi_i(0)$ stands for the proportion of base i at the time when divergence starts. We can think of $X(0)$ as the result of a comparison of the ancestral nucleotide sequence with itself and hence the diagonal character of $X(0)$.

The matrix $X(t)$ is symmetric provided it is $X(0)$, for real H. In that case, $\bar{P}_i = P_i, \bar{Q}_i = Q_i, \bar{R}_i = R_i$, $(i = 1, 2)$; and we are left with just ten functions.

Evolutionary Matrix (or Time–Evolution Matrix): $U(t) = (u_{ij}(t))$, $i, j = 1, \ldots, 4$. This matrix involves one nucleotide sequence. Given a site, $u_{ij}(t)$ stands for the conditional probability that there is a base i at time t when at time $t = 0$ there was a base j. Thus $U(0) = I$ (identity matrix). The sum of the matrix elements of every single column of $U(t)$ equals unity.

The time–evolution matrix answers the question: Given the base proportions in the nucleotide sequence at $t = 0$, what are their values at time t? The algebraic solution is readily written down if we use the vector whose components are the base proportions at time t: $(\pi_T(t), \pi_C(t), \pi_A(t), \pi_G(t))^\intercal$, where the superscript stands for the transposed matrix. At time t we get

$$\begin{pmatrix} \pi_T(t) \\ \pi_C(t) \\ \pi_A(t) \\ \pi_G(t) \end{pmatrix} = U(t) \begin{pmatrix} \pi_T(0) \\ \pi_C(0) \\ \pi_A(0) \\ \pi_G(0) \end{pmatrix}. \tag{4}$$

As we will see below, $U(t) = \exp(Ht)$. The exponential of a square matrix is defined by the Taylor expansion

$$\exp(Ht) = I + \frac{1}{1!}Ht + \frac{1}{2!}H^2t^2 + \frac{1}{3!}H^3t^3 + \ldots, \tag{5}$$

and is a square matrix with the dimensions of H. At first, all the computational burden reduces to evaluate powers of the matrix H. Fortunately, a number of algebraic techniques exist to carry out the task efficiently [13, 14].

For Kimura rates matrix (1), henceforth denoted H_k, it is not difficult to obtain the closed form expression

$$\exp(H_k t) = \tag{6}$$
$$\frac{1}{4}\begin{pmatrix} 1+\lambda_{\alpha\beta}+\lambda_{\alpha\gamma}+\lambda_{\beta\gamma} & 1-\lambda_{\alpha\beta}-\lambda_{\alpha\gamma}+\lambda_{\beta\gamma} & 1-\lambda_{\alpha\beta}+\lambda_{\alpha\gamma}-\lambda_{\beta\gamma} & 1+\lambda_{\alpha\beta}-\lambda_{\alpha\gamma}-\lambda_{\beta\gamma} \\ 1-\lambda_{\alpha\beta}-\lambda_{\alpha\gamma}+\lambda_{\beta\gamma} & 1+\lambda_{\alpha\beta}+\lambda_{\alpha\gamma}+\lambda_{\beta\gamma} & 1+\lambda_{\alpha\beta}-\lambda_{\alpha\gamma}-\lambda_{\beta\gamma} & 1-\lambda_{\alpha\beta}+\lambda_{\alpha\gamma}-\lambda_{\beta\gamma} \\ 1-\lambda_{\alpha\beta}+\lambda_{\alpha\gamma}-\lambda_{\beta\gamma} & 1+\lambda_{\alpha\beta}-\lambda_{\alpha\gamma}-\lambda_{\beta\gamma} & 1+\lambda_{\alpha\beta}+\lambda_{\alpha\gamma}+\lambda_{\beta\gamma} & 1-\lambda_{\alpha\beta}-\lambda_{\alpha\gamma}+\lambda_{\beta\gamma} \\ 1+\lambda_{\alpha\beta}-\lambda_{\alpha\gamma}-\lambda_{\beta\gamma} & 1-\lambda_{\alpha\beta}+\lambda_{\alpha\gamma}-\lambda_{\beta\gamma} & 1-\lambda_{\alpha\beta}-\lambda_{\alpha\gamma}+\lambda_{\beta\gamma} & 1+\lambda_{\alpha\beta}+\lambda_{\alpha\gamma}+\lambda_{\beta\gamma} \end{pmatrix},$$

where, for the sake of readability, we have defined the quantities

$$\begin{aligned} \lambda_{\alpha\beta} &= \exp[-2(\alpha+\beta)t], \\ \lambda_{\alpha\gamma} &= \exp[-2(\alpha+\gamma)t], \\ \lambda_{\beta\gamma} &= \exp[-2(\beta+\gamma)t]. \end{aligned} \tag{7}$$

In the case of JC matrix of substitution rates, denoted H_{jc}, we have $\alpha = \beta = \gamma \equiv \mu$ and so the result above collapses to

$$\exp(H_{jc}t) = \begin{pmatrix} \delta & \xi & \xi & \xi \\ \xi & \delta & \xi & \xi \\ \xi & \xi & \delta & \xi \\ \xi & \xi & \xi & \delta \end{pmatrix}, \quad \delta = \frac{1}{4}(1+3\mathrm{e}^{-4\mu t}), \quad \xi = \frac{1}{4}(1-\mathrm{e}^{-4\mu t}). \tag{8}$$

3. Time–evolution equation

Given a substitution model, defined by H, time–evolution is determined by the matrix product

$$X(t) = U(t)\,X(0)\,U^{\mathsf{T}}(t), \tag{9}$$

where the matrix $U(t)$ is the solution of the matrix differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}U(t) = HU(t), \qquad U(0) = I. \tag{10}$$

Whenever H is a constant matrix the solution reads

$$U(t) = \exp(Ht), \qquad U^{\mathsf{T}}(t) = \exp(H^{\mathsf{T}}t). \tag{11}$$

Thus,

$$X(t) = \mathrm{e}^{Ht}X(0)\,\mathrm{e}^{H^{\mathsf{T}}t}. \tag{12}$$

The initial condition $X(0)$ is taken as the pure diagonal matrix (3). In addition, the divergence matrix obeys its own differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}X(t) = HX(t) + X(t)H^{\mathsf{T}}. \tag{13}$$

Eventually, once base proportions are fixed at $t = 0$, the time–evolution of the solution is fully determined and equation (12) provides the connection between the inputs of the mathematical model and the experimental measurements obtained with two homologous sequences.

4. JC and K3ST substitution models solved in matrix formalism

The analytic solution for the matrix elements of $X(t)$ is readily obtained after carrying out the matrix products in (12). Due to the symmetry of matrix H_k (see also the scheme in Figure 1) the equations for $\bar{P}_i, \bar{Q}_i, \bar{R}_i$, $(i = 1, 2)$ in (2) are formally identical to those for P_i, Q_i, R_i, $(i = 1, 2)$. This symmetry can be exploited to reduce the dimension of the system of differential equations from sixteen to ten. The idea is to consider altogether substitutions $i \to j$ and $j \to i$. Namely, we study the content AT+TA, CT+TC, GT+TG, AG+GA, CG+GC, and AC+CA, and thus base pairs ij and ji are indistinguishable configurations when comparing the base content of homologous sequences. The way to implement this is via a dimensional reduction considering new probability functions: $p_i \equiv P_i + \bar{P}_i$, $q_i \equiv Q_i + \bar{Q}_i$, $r_i \equiv R_i + \bar{R}_i$, $i = 1, 2$; and, for the sake of notational convenience, $s_i \equiv S_i$, $i = 1$ to 4, too. The explicit solutions appear in the central column of Table 2. We have defined $\Gamma = \frac{1}{16}\left(1 + \lambda^2_{\alpha\beta} + \lambda^2_{\alpha\gamma} + \lambda^2_{\beta\gamma}\right)$. The last column applies only with initial condition (31) below to include ancestral polymorphisms. The four initial (constrained) proportions have been expressed in terms of three free parameters, σ_p, σ_q, and σ_r, defined as follows

$$\begin{aligned}
\sigma_p &= \pi_A(0) + \pi_G(0) - \pi_C(0) - \pi_T(0), \\
\sigma_q &= \pi_C(0) + \pi_G(0) - \pi_A(0) - \pi_T(0), \\
\sigma_r &= \pi_G(0) + \pi_T(0) - \pi_A(0) - \pi_C(0). \qquad (14)
\end{aligned}$$

The inverse transformation reads

$$\begin{aligned}
\pi_A(0) &= \frac{1}{4}(1 + \sigma_p - \sigma_q - \sigma_r), \quad \pi_C(0) = \frac{1}{4}(1 - \sigma_p + \sigma_q - \sigma_r), \\
\pi_G(0) &= \frac{1}{4}(1 + \sigma_p + \sigma_q + \sigma_r), \quad \pi_T(0) = \frac{1}{4}(1 - \sigma_p - \sigma_q + \sigma_r). \qquad (15)
\end{aligned}$$

The numerical combinations of parameters $\sigma_p, \sigma_q, \sigma_r$, that fulfil the constraints $0 \le \pi_i(0) \le 1$, fill in the volume of a tetrahedron with vertices $(1, 1, 1)$, $(1, -1, -1)$, $(-1, 1, -1)$ and $(-1, -1, 1)$ in the three–dimensional σ–parameter space. The use of these parameters has been instrumental in shortening the formulas of time–dependent solutions and no further interpretation is intended.

It is worth reminding that the initial base proportions are, at first, arbitrary. Hence, when analysing two homologous sequences the number of free parameters in the mathematical scheme is: three ancestral sequence base proportions at $t = 0$ plus the number of free parameters in the rates matrix H. For JC and K3ST schemes this number is $3 + 1$ and $3 + 3$ respectively.

The results in Table 2 are pre–K3ST solutions in the sense that they do not match those reported by Kimura [5] which are obtained after carrying out the following phase space dimensional reduction

$$S = s_1 + s_2 + s_3 + s_4, \quad P = p_1 + p_2, \quad Q = q_1 + q_2, \quad R = r_1 + r_2. \qquad (16)$$

This notation is not to be confused with that in (2) where P, Q, R, S, symbols are subindexed.

The pre–K3ST solution may be interpreted as a trajectory that evolves with time in a phase space with ten coordinate axes: $\{s_1, s_2, s_3, s_4, p_1, p_2, q_1, q_2, r_1, r_2\}$. Then we project it onto a lower dimension phase space with only four axes: $\{S, P, Q, R\}$, according to the transformation (16), yielding Kimura phase space trajectory. Thus, summing up the corresponding equations in Table 2 we get the very result reported by Kimura

$$\begin{aligned}
S(t) &= \frac{1}{4}\left(1 + \lambda^2_{\alpha\beta} + \lambda^2_{\alpha\gamma} + \lambda^2_{\beta\gamma}\right), \\
P(t) &= \frac{1}{4}(1 - \lambda^2_{\alpha\beta} - \lambda^2_{\alpha\gamma} + \lambda^2_{\beta\gamma}), \\
Q(t) &= \frac{1}{4}(1 - \lambda^2_{\alpha\beta} + \lambda^2_{\alpha\gamma} - \lambda^2_{\beta\gamma}), \\
R(t) &= \frac{1}{4}(1 + \lambda^2_{\alpha\beta} - \lambda^2_{\alpha\gamma} - \lambda^2_{\beta\gamma}), \qquad (17)
\end{aligned}$$

with λ_{ij} given in (7). It can be readily checked that the sum of these four functions is equal to one. Notice that no explicit dependence on the initial base proportions appears, and the initial values here are fixed: $S(0) = 1, P(0) = Q(0) = R(0) = 0$. The independence of this solution with respect to the initial base proportions $\pi_i(0)$, is a mere consequence of the dimensional reduction in (16). The essence of this approach consists in discarding part of the information of the experimental measurements of nucleotide substitutions by combining them according to the scheme: TT+CC+AA+GG, CT+TC+AG+GA, AT+TA+CG+GC, GT+TG+AC+CA. The advantage of this procedure is that the reduced system of differential equations is much easier to solve.

Eventually, if in Table 2 we set $\alpha = \beta = \gamma \equiv \mu$, the general solution for the pre–JC model is readily obtained (see Table 3). The conventional solution for the JC model emerges after the dimensional reduction

$$M = p_1 + p_2 + q_1 + q_2 + r_1 + r_2, \quad S = s_1 + s_2 + s_3 + s_4. \tag{18}$$

The final result reads

$$\begin{aligned} S(t) &= \frac{1}{4}\left[1 + 3\exp(-8\mu t)\right], \\ M(t) &= \frac{3}{4}\left[1 - \exp(-8\mu t)\right]. \end{aligned} \tag{19}$$

Solution (19) is independent of the initial base proportions and the same considerations as above apply. In this case, only information about the total number of substitutions between homologous sequences is retained.

Finally, we write explicitly the components of the vector of base proportions of a single nucleotide sequence as a function of time obtained with (4)

$$\begin{aligned} 4\pi_T(t) &= 1 + \sigma_r\,\lambda_{\alpha\beta} - \sigma_q\,\lambda_{\alpha\gamma} - \sigma_p\,\lambda_{\beta\gamma}, \\ 4\pi_C(t) &= 1 - \sigma_r\,\lambda_{\alpha\beta} + \sigma_q\,\lambda_{\alpha\gamma} - \sigma_p\,\lambda_{\beta\gamma}, \\ 4\pi_A(t) &= 1 - \sigma_r\,\lambda_{\alpha\beta} - \sigma_q\,\lambda_{\alpha\gamma} + \sigma_p\,\lambda_{\beta\gamma}, \\ 4\pi_G(t) &= 1 + \sigma_r\,\lambda_{\alpha\beta} + \sigma_q\,\lambda_{\alpha\gamma} + \sigma_p\,\lambda_{\beta\gamma}, \end{aligned} \tag{20}$$

for K3ST model, and simply

$$\pi_i(t) = \frac{1}{4} + \left(\pi_i(0) - \frac{1}{4}\right)\exp(-4\mu t), \qquad i : T, C, A, G, \tag{21}$$

for JC model. Notice the explicit dependence of the solution on ancestral proportions. It can be readily checked that the sum of the four $\pi_i(t)$ is equal to one in (20) and (21). These four functions describe a trajectory in a phase space of four dimensions which is unique once the ancestral proportions are given. The dimension of this phase space coincides with Kimura's. It is then convenient to point out that they are phase spaces of different nature. The trajectories described by $\{\pi_T(t), \pi_C(t), \pi_A(t), \pi_G(t)\}$ and $\{S(t), P(t), Q(t), R(t)\}$, do not admit comparison at all.

It is customary to assume that the ancestral population is the result of a long enough time–evolution so that an *equilibrium* regime has been reached before the bifurcation takes place. Equilibrium means here that the proportions $\pi_i(t)$ remain constant in time. This assumption is consistent with the mathematical asymptotic behaviour of solutions (20) and (21) because for large t all the exponential terms vanish and we are left with the constant vector of proportions $\frac{1}{4}(1,1,1,1)^\intercal$. Equilibrium with equal base proportions is a consequence of the symmetry of the matrices H_k and H_{jc}. Equilibrium with unequal base proportions are (essentially) associated to non symmetric H matrices. It can be shown that equilibrium proportions are determined by the coordinates of the eigenvector $\vec{v}$ of H with vanishing eigenvalue, i.e. $H\vec{v} = 0$.

When time–evolution from equilibrium is assumed, namely $\pi_A(0) = \pi_C(0) = \pi_G(0) = \pi_T(0) = \frac{1}{4}$, then $\sigma_p = \sigma_q = \sigma_r = 0$, and the full ten dimensional solutions of K3ST model in Table 2 simplify to those in Table 4.

5. Evolutionary distance with respect to the ancestor

As stated in Section 2, the time–evolution matrix $U(t)$ gives the new proportions of the ancestral base sequence after a time t has elapsed. The likely number of substitutions to occur at instant t is given by $\sum_{ij,i\neq j} \pi_i(t)\, H_{ji} = -\sum_i \pi_i(t)\, H_{ii}$. The average number of substitutions ν accumulated since $t = 0$ up to time t is obtained by integration

$$\nu = -2 \sum_{i=1}^{4} \int_0^t H_{ii}\, \pi_i(t')\, \mathrm{d}t', \tag{22}$$

where the additional factor two implies that this distance is between the tips of two lineages. Usually, the H_{ii} terms are time independent and can be taken out from the integral. This is a definition for evolutionary distance and, unless steady time–evolution is assumed, explicit knowledge of $\pi_i(t)$, $i = 1, \ldots, 4$, is needed. This, in turn, involves the values of the initial base proportions as well which is a true hindrance because they are unknown. It is here where the assumption of steady time–evolution helps and is on the basis of the so–called *General Time Reversible* models [15]. Non–stationary time–evolution is certainly a difficult problem to which attention has been paid in the literature [4, 16]. Fortunate enough, it turns out that the evolutionary distances ν_{jc} and ν_k, obtained from (22) with (21) and (20), are independent of the ancestral base proportions

$$\nu_{jc} = 6\mu t, \tag{23}$$

$$\nu_k = 2(\alpha + \beta + \gamma)t, \tag{24}$$

due to the high symmetry of the matrices H_{jc} and H_k. Therefore, no assumption about equilibrium is required. These formulas provide the formal evolutionary distance between two homologous sequences in terms of the model parameters and the time variable. Practical evolutionary distance formulas appear once a connection between $\alpha t, \beta t, \gamma t, \mu t$ and the matrix elements of $X(t)$, which contain the experimental measurements, is established.

Let us denote with a tilde all experimentally measured value. Thus, $\tilde{M}$ stands for the experimental determination of the total mutation probability $M(t)$, estimated from comparison of the two homologous nucleotide sequences at hand. Then, the JC model equation (19) yields

$$\exp(-8\mu t_{jc}) = 1 - \frac{4}{3}\tilde{M}, \tag{25}$$

and, after some algebra, equations (17) lead to

$$\begin{aligned}
\exp(-8\alpha t_k) &= \frac{[1 - 2(\tilde{P} + \tilde{Q})][1 - 2(\tilde{P} + \tilde{R})]}{1 - 2(\tilde{Q} + \tilde{R})}, \\
\exp(-8\beta t_k) &= \frac{[1 - 2(\tilde{P} + \tilde{Q})][1 - 2(\tilde{Q} + \tilde{R})]}{1 - 2(\tilde{P} + \tilde{R})}, \\
\exp(-8\gamma t_k) &= \frac{[1 - 2(\tilde{P} + \tilde{R})][1 - 2(\tilde{Q} + \tilde{R})]}{1 - 2(\tilde{P} + \tilde{Q})},
\end{aligned} \tag{26}$$

with $\tilde{P}, \tilde{Q}, \tilde{R}$, estimated by the comparison of the two homologous nucleotide sequences too. Substitution of (25) in (23) gives the classical JC approximation for the number of mutations *per site* occurred since divergence

$$\nu_{jc} = -\frac{3}{4} \ln\left(1 - \frac{4}{3}\tilde{M}\right), \tag{27}$$

provided $\tilde{M} < 3/4$. Substitution of (26) in (24) leads to the well known Kimura's formula

$$\nu_k = -\frac{1}{4} \ln\left\{[1 - 2(\tilde{P} + \tilde{Q})][1 - 2(\tilde{P} + \tilde{R})][1 - 2(\tilde{Q} + \tilde{R})]\right\}, \tag{28}$$

provided the three square bracket are positive–valued. These two equations establish approximate determinations of the evolutionary distance in terms of experimental measurements for the substitution frequencies of the sequences of both lineages.

The algebraical procedure followed has allowed a proper derivation of the classical evolutionary distance associated to both JC and K3ST models. We have stressed the fact that none of the results are flawed by the steady time–evolution assumption. Next we show that the matrix scheme proves particularly useful to incorporate ancestral polymorphisms and to deplete the effect of deleterious mutations.

Eventually, measurements $\tilde{M}, \tilde{P}, \tilde{Q}, \tilde{R}$, are carried out with nucleotide sequences of finite length n, a fact that conveys some degree of indetermination. The large sampling variance of the evolutionary distance has been estimated as

$$\sigma^2_{jc} = \frac{\tilde{M}(1 - \tilde{M})}{n\left(1 - \frac{4}{3}\tilde{M}\right)^2}, \tag{29}$$

for JC model [17] and

$$\begin{aligned} \sigma^2_k &= \frac{1}{4n}\left[a^2\tilde{P} + b^2\tilde{Q} + c^2\tilde{R} - (a\tilde{P} + b\tilde{Q} + c\tilde{R})^2\right] \\ a &= \frac{1}{1 - 2\tilde{P} - 2\tilde{Q}} + \frac{1}{1 - 2\tilde{P} - 2\tilde{R}}, \\ b &= \frac{1}{1 - 2\tilde{P} - 2\tilde{Q}} + \frac{1}{1 - 2\tilde{Q} - 2\tilde{R}}, \\ c &= \frac{1}{1 - 2\tilde{P} - 2\tilde{R}} + \frac{1}{1 - 2\tilde{Q} - 2\tilde{R}}, \end{aligned} \tag{30}$$

for K3ST model [5]. These estimates can be also used with the generalized formulas bellow.

6. Evolutionary distance and ancestral polymorphisms

The rationale to incorporate ancestral polymorphisms into the molecular clock is the following. As pointed out in item 2 of Section 2, $X(t = 0)$ can be thought as the result of a comparison of the ancestral nucleotide sequence with itself. If the ancestral population presents polymorphisms it is then natural to consider a non–diagonal $X(0)$ matrix as an effective way to incorporate this effect

$$X(t = 0) = \begin{pmatrix} \pi_T(0) - 3\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & \pi_C(0) - 3\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & \pi_A(0) - 3\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & \pi_G(0) - 3\epsilon \end{pmatrix}, \tag{31}$$

with $\pi_A(0) + \pi_C(0) + \pi_G(0) + \pi_T(0) = 1$, so as to preserve probability and $0 \leq \epsilon \leq \min\{\pi_T, \pi_C, \pi_A, \pi_G\}$. The parameter ϵ is intended to account for the probability of different base pairs in a random drawn of two individuals from the ancestral population and expected $\epsilon \ll 1$. The idea is to build up evolutionary distance formulas that remove the effect of those base mutations belonging to a different dynamical process (namely, heterozygosity of the ancestral population).

A straightforward calculation with (31), (12) and (6) leads to the analytical solutions for $X(t)$ in Table 3, where the rightmost column has now to be included in the solution. Proceeding as above we obtain the JC–like evolutionary distance that includes the effect of ancestral polymorphisms

$$6\mu t_{jcp} = \nu_{jcp} = -\frac{3}{4}\left[\ln\left(1 - \frac{4}{3}M\right) - \ln\left(1 - 16\epsilon\right)\right], \tag{32}$$

provided the right hand side is non–negative–valued, which occurs whenever $0 < 12\epsilon < M < 3/4$. The constraint $12\epsilon < M$ ensures that the amount of mutations at modern time is greater than the ancestral polymorphisms. The subscript in ν_{jcp} refers to JC model with ancestral polymorphisms. Interestingly, this formula has been obtained on the basis of probability arguments in [8], where 12ϵ is to be interpreted as the *expected heterozygosity*. This effect does not explain fully the apparent mutation rate acceleration observed at short evolution times but, indeed, provides an interesting partial answer to the phenomenon. In human populations, heterozygosity is around 0.001 [18], indeed much smaller than $\pi_i(0)$.

When we consider the rather general parametrization of the initial divergence matrix

$$X(0) = \begin{pmatrix} \pi_T(0) - \eta - \rho - \varphi & \eta & \rho & \varphi \\ \eta & \pi_C(0) - \eta - \sigma - \xi & \sigma & \xi \\ \rho & \sigma & \pi_A(0) - \rho - \sigma - \nu & \zeta \\ \varphi & \xi & \zeta & \pi_G(0) - \varphi - \xi - \zeta \end{pmatrix}, \quad (33)$$

in which every type of polymorphism has different weight $\eta, \rho, \varphi, \sigma, \xi$, and ζ, all assumed to be small enough and positive; the calculation leads to the JC extended formula

$$\nu_{jcp+} = -\frac{3}{4}\left[\ln\left(1 - \frac{4}{3}M\right) - \ln\left(1 - \frac{8}{3}(\eta + \rho + \varphi + \sigma + \xi + \zeta)\right)\right], \quad (34)$$

provided the bracket is non–negative–valued. The expected heterozygosity is here $2(\eta + \rho + \varphi + \sigma + \xi + \zeta)$, i.e the fraction of polymorphisms assumed in the ancestral population. Notice that ν_{jcp+} does not depend on partial heterozygosities but on the sum of all of them.

For the sake of simplicity, in the following generalization to K3ST model we consider only the one parameter ansatz (31). The analytical solution for $X(t)$ appears in Table 2, where the rightmost column has to be included now. The final expression for the K3ST evolutionary distance with ancestral polymorphisms reads

$$\nu_{kp} = -\frac{1}{4}\Big(\ln\{[1 - 2(P + Q)][1 - 2(P + R)][1 - 2(Q + R)]\} - 3\ln(1 - 16\epsilon)\Big), \quad (35)$$

provided the right hand side is non–negative–valued.

7. Evolutionary distance and purifying selection

An interesting explanation for the apparent acceleration of mutation rates at short time scales has been developed in terms of the presence of deleterious mutations [9, 10]. Correcting for purifying selection introduces a time dependency in the substitution rates matrix. Thus, the constant matrix of substitution rates H is replaced with the product $Hf(t)$, where $f(t)$ is an appropriate decreasing (usually described as $J-$shaped) scalar function of time when plotted from present to past. An exponentially decaying profile has been proposed although, most likely, the particular shape details are not crucial. Here, we adhere to this choice and solve the time–dependent system in matrix formalism to obtain the modified JC and K3ST evolutionary distances in which the effect of deleterious mutations has been removed as far as the scheme is correct.

We commence by noticing that the time–evolution matrix equation (10) with the time–dependent rate matrix $Hf(t)$, is still analytically solvable. The mathematical reason is that the two time–dependent matrices $Hf(t)$ and $\int Hf(t')\mathrm{d}t'$, commute with respect to the matrix product for all values of t [19].

The crucial modification in the following consists in integrating the evolution equations backward in time as already alluded to in the Introduction. This is because $f(t)$ is well defined from present to past which fixes the origin of coordinate t. Thus, $t = 0$ refers to modern time. The integration progresses towards negative time values until we eventually meet at $-t$ (with $0 < t$) a coalescence constraint. It can be shown [19] that under these conditions, the time–evolution matrix that solves the problem is

$$U(t) = \exp\left[HF(t)\right], \qquad F(t) = \int_0^{-t} f(t')\mathrm{d}t', \qquad 0 < t. \quad (36)$$

The exponential matrix is obtained from (6) simply replacing t with $F(t)$. The exponential ansatz $f(t) = 1 + a\,\exp(t/\tau)$, with $t < 0$, introduces two new parameters, namely, the purifying selection time scale $\tau > 0$, and the dimensionless amplitude $a \geq 0$ [2, 9, 10, 20]. For the present purposes we assume that a and τ have already been estimated. The integral in (36) gives

$$F(t) = -\Big(t + a\tau[1 - \exp(-t/\tau)]\Big), \qquad 0 < t. \tag{37}$$

Backward time integration of molecular clock equations requires to take as initial condition the measured values of the divergence matrix

$$X(0) = \begin{pmatrix} \tilde{s}_1 & \tilde{p}_1 & \tilde{q}_1 & \tilde{r}_1 \\ \tilde{p}_1 & \tilde{s}_2 & \tilde{r}_2 & \tilde{q}_2 \\ \tilde{q}_1 & \tilde{r}_2 & \tilde{s}_3 & \tilde{p}_2 \\ \tilde{r}_1 & \tilde{q}_2 & \tilde{p}_2 & \tilde{s}_4 \end{pmatrix}, \tag{38}$$

and evolve backward in time with the matrix product

$$X(t) = \mathrm{e}^{HF(t)}\, X(0)\, \mathrm{e}^{H^{\intercal}F(t)}. \tag{39}$$

This matrix product yields the pre–Kimura solution of dimension ten. The dimensional reductions (16) and (18) provide the solutions for K3SP and JC substitution models respectively. Eventually we impose the algebraic constraint(s) that define coalescence. Strictly speaking, that time has been reached when the divergence matrix $X(t)$ becomes diagonal as in (3) which conveys up to six constraints on the non–diagonal matrix elements of X. However, JC is a one–parameter model and hence only one constraint is viable: $M(t) \equiv \sum_{i\neq j} X_{ij}(t) = 0$. For K3ST model we have three: $P(t) \equiv X_{12}(t) + X_{21}(t) + X_{34}(t) + X_{43}(t) = 0$, $Q(t) \equiv X_{13}(t) + X_{31}(t) + X_{24}(t) + X_{42}(t) = 0$, and $R(t) \equiv X_{14}(t) + X_{41}(t) + X_{23}(t) + X_{32}(t) = 0$. A straightforward calculation leads to

$$\exp(8\mu F(t)) = 1 - \frac{4}{3}\tilde{M}, \tag{40}$$

and

$$\begin{aligned} \exp(8\alpha F(t)) &= \frac{[1 - 2(\tilde{P} + \tilde{Q})][1 - 2(\tilde{P} + \tilde{R})]}{1 - 2(\tilde{Q} + \tilde{R})}, \\ \exp(8\beta F(t)) &= \frac{[1 - 2(\tilde{P} + \tilde{Q})][1 - 2(\tilde{Q} + \tilde{R})]}{1 - 2(\tilde{P} + \tilde{R})}, \\ \exp(8\gamma F(t)) &= \frac{[1 - 2(\tilde{P} + \tilde{R})][1 - 2(\tilde{Q} + \tilde{R})]}{1 - 2(\tilde{P} + \tilde{Q})}, \end{aligned} \tag{41}$$

for JC and K3ST models respectively. Notice the similarity with (25) and (26).

From (40) and (37) we obtain

$$-8\mu\,\{t + a\tau[1 - \exp(-t/\tau)]\} = \ln\left(1 - \frac{4}{3}\tilde{M}\right), \tag{42}$$

to be solved for $t > 0$, which stands for the coalescence time measured from present ($t = 0$). Similarly, proceeding with (42) we obtain for K3ST model

$$\begin{aligned} &-8(\alpha + \beta + \gamma)\,\{t + a\tau[1 - \exp(-t/\tau)]\} = \\ &\qquad \ln\left\{[1 - 2(\tilde{P} + \tilde{Q})][1 - 2(\tilde{P} + \tilde{R})][1 - 2(\tilde{Q} + \tilde{R})]\right\}. \end{aligned} \tag{43}$$

These formulas determine implicitly the coalescence time in terms of experimental estimates of the substitution probabilities. It is easy to check that if we put $a = 0$ in (42) and

(43), the conventional JC and K3ST results are recovered. This is a verification about the equivalence between forward and backward time integration formalisms.

The evolutionary distance along lineages is computed just as in Section 5, namely, using the constant rates matrix H, and not $Hf(t)$. The effective time–varying rates matrix is only used when comparing the modern homologous molecular sequences because we are unable to disentangle among neutral and deleterious mutations in the measurement process. Let us introduce the new symbols $\bar{\nu}_{jc} = 6\mu t$ and $\bar{\nu}_k = 2(\alpha + \beta + \gamma)t$, for the evolutionary distances free from deleterious mutations to mean that the coalescence time determined will differ from the classical JC and K3ST formulas. For the sake of convenience we define the two evolutionary scales $\nu_\tau^{jc} = 6\mu\tau$ and $\nu_\tau^k = 2(\alpha + \beta + \gamma)\tau$, too. Equations (42) and (43) can be now written down in a unified way

$$\frac{\nu_s}{\nu_\tau^s} = \frac{\bar{\nu}_s}{\nu_\tau^s} + a\left[1 - \exp\left(-\frac{\bar{\nu}_s}{\nu_\tau^s}\right)\right], \qquad s : jc,\, k. \tag{44}$$

These formulas connect the corrected for purifying selection evolutionary distances $\bar{\nu}_s$ and the uncorrected ν_s, in units of their own proper purifying selection scales ν_τ^s. Moreover, (44) shows that $\bar{\nu}_s < \nu_s$, for both models. Equation (44) provides only *implicit* solutions for $\bar{\nu}_s$. *Explicit* solutions are given in next section.

8. Ancestral polymorphisms and purifying selection combined

To combine the effect of ancestral polymorphisms and purifying selection the only technical modification with respect to Section 7 concerns the algebraic constraints that in backward integration defines coalescence in presence of heterozygosity. Or so to speak: *imperfect coalescence*. For JC model the only constraint is now $M(t) \equiv \sum_{i \neq j} X_{ij}(t) = 12\epsilon$; and for K3ST model: $P(t) \equiv X_{12}(t) + X_{21}(t) + X_{34}(t) + X_{43}(t) = 4\epsilon$, $Q(t) \equiv X_{13}(t) + X_{31}(t) + X_{24}(t) + X_{42}(t) = 4\epsilon$, and $R(t) \equiv X_{14}(t) + X_{41}(t) + X_{23}(t) + X_{32}(t) = 4\epsilon$. After straightforward calculation we find the very same formal equation (44) extended to embrace ν_{jcp} and ν_{kp}. Moreover, despite the non–linear character of (44) the solutions may be explicitly written down in terms of Lambert W function[1] [21]

$$\frac{\bar{\nu}_s}{\nu_\tau^s} = \frac{\nu_s}{\nu_\tau^s} - a + W\left(a \exp\left(a - \frac{\nu_s}{\nu_\tau^s}\right)\right), \quad \nu_\tau^s = \begin{cases} 6\mu\tau, & s : jc,\, jcp, \\ 2(\alpha+\beta+\gamma)\tau, & s : k,\, kp. \end{cases} \tag{45}$$

This result generalizes the classical JC and K3ST formulas to incorporate the effect of purifying selection modelled via an exponential decaying function to deplete the effect of deleterious mutations. It applies to JC and K3ST evolutionary distances with and without ancestral polymorphisms.

The non–linear character of the relationship between $\bar{\nu}_s$ and ν_s is witnessed by the series expansion

$$\frac{\bar{\nu}_s}{\nu_\tau^s} \simeq \frac{\nu_s}{\nu_\tau^s} - a\left[1 - \exp\left(-\frac{\nu_s}{\nu_\tau^s}\right)\right]\left[1 - a\exp\left(-\frac{\nu_s}{\nu_\tau^s}\right)\right] + \mathcal{O}(a^3). \tag{46}$$

This second order approximation avoids the explicit computation of the Lambert W function.

[1] Lambert W function satisfies $W(x)\exp(W(x)) = x$, and provides the solution of the transcendental equation $x + A + B\exp(Cx) = 0$, as $x = -A - \frac{1}{C}W(BC\exp(-AC))$. Numerical routines exist in common programming languages.

9. Backward in time: *Ab initio* removal of deleterious mutations

Let us suppose we are able to estimate the amount of deleterious mutations present in the two homologous nucleotide sequences, just in the same way as we did in Section 6 with ancestral polymorphisms. For the sake of simplicity, we will use a sole parameter, say δ, to quantify the effect.

The proposal is to take the divergence matrix at present time as initial condition

$$X(0) = \begin{pmatrix} \tilde{s}_1 + 3\delta & \tilde{p}_1 - \delta & \tilde{q}_1 - \delta & \tilde{r}_1 - \delta \\ \tilde{p}_1 - \delta & \tilde{s}_2 + 3\delta & \tilde{r}_2 - \delta & \tilde{q}_2 - \delta \\ \tilde{q}_1 - \delta & \tilde{r}_2 - \delta & \tilde{s}_3 + 3\delta & \tilde{p}_2 - \delta \\ \tilde{r}_1 - \delta & \tilde{q}_2 - \delta & \tilde{p}_2 - \delta & \tilde{s}_4 + 3\delta \end{pmatrix}, \tag{47}$$

and to integrate backward in time till coalescence. Remember that all elements with tilde correspond to values determined experimentally, as in (38). The value of δ with $0 \le \delta \le \min\{\tilde{p}_i, \tilde{q}_i, \tilde{r}_i\}$, $i = 1, 2$, is assumed to be known and stands for an estimate of deleterious mutation fraction embedded in the two homologous nucleotide sequence samples. The rationale is that, proceeding this way, deleterious mutations have been successfully removed *ab initio* from the divergence matrix so that backward time integration takes place with the constant matrix H. Therefore equation (39) simply reads now

$$X(t) = \mathrm{e}^{-Ht} X(0) \mathrm{e}^{-H^{\top} t}, \tag{48}$$

and t refers to past. An analogous algebraic computation as in previous sections leads to the simple formulas

$$6\mu t \equiv \breve{\nu}_{jc} = -\frac{3}{4} \ln\left[1 - \frac{4}{3}(\tilde{M} - 12\delta)\right], \tag{49}$$

and

$$2(\alpha+\beta + \gamma)t \equiv \breve{\nu}_k = \\ -\frac{1}{4} \ln\left\{[1 - 2(\tilde{P} + \tilde{Q} - 8\delta)][1 - 2(\tilde{P} + \tilde{R} - 8\delta)][1 - 2(\tilde{Q} + \tilde{R} - 8\delta)]\right\}, \tag{50}$$

for JC and K3ST models respectively. Here, $\breve{\nu}_i$, with $i : jc, k$, refers to evolutionary distance with initial deleterious mutations removed prior to backward time integration. The interpretation for both formulas is: Use either JC or K3ST distance with deleterious mutation fraction δ removed from the measurable quantities; namely, map either $\tilde{M} \to \tilde{M} - 12\delta$, or $\tilde{P} \to \tilde{P} - 4\delta$, $\tilde{Q} \to \tilde{Q} - 4\delta$, and $\tilde{R} \to \tilde{R} - 4\delta$, respectively. Recall that each $\tilde{P}, \tilde{Q}, \tilde{R}$, is a sum of four non–diagonal divergence matrix elements, as a consequence of the phase space dimensional reduction and hence the term 4δ in the mapping.

10. Ancestral polymorphisms and *ab initio* removal of deleterious mutations

A heuristic generalization of (49) to include ancestral polymorphisms goes as follows. Equation (19) is the implicit JC form relating the mutation fraction M and the time elapsed t. Let us interpret ancestral heterozygosity as an *imperfect coalescence* scenario in which a small mutation fraction $12\epsilon \equiv M_0 > 0$, is present. The sequences must have diverged t_0 time units in the past. Let us assume that both quantities are related by JC evolutionary distance formula. Therefore, the equation $6\mu t_0 = -\frac{3}{4}\ln(1 - \frac{4}{3}M_0)$, maps M_0 onto t_0, and viceversa. Next let us assume that a fraction $12\delta \equiv M_d$ out of M corresponds to deleterious mutations. Then the *true* neutral mutation fraction and evolutionary time are respectively: $M \to M - M_d$ and $t \to t + t_0$ in (19). Thus

$$M - M_d = \frac{3}{4}\{1 - \exp[-8\mu(t + t_0)]\}. \tag{51}$$

After algebraic inversion

$$6\mu(t+t_0) = -\frac{3}{4}\ln\left[1-\frac{4}{3}(M-M_d)\right], \tag{52}$$

or

$$6\mu t = -\frac{3}{4}\left\{\ln\left[1-\frac{4}{3}(M-12\delta)\right]-\ln\left(1-16\epsilon\right)\right\}. \tag{53}$$

Ancestral heterozygosity and deleterious mutation removal corrections act as time origin and mutation fraction shifts, respectively.

The generalization for K3ST reads

$$\begin{aligned}&2(\alpha+\beta+\gamma)t = \\ &-\frac{1}{4}\left(\ln\left\{[1-2(\tilde{P}+\tilde{Q}-8\delta)][1-2(\tilde{P}+\tilde{R}-8\delta)][1-2(\tilde{Q}+\tilde{R}-8\delta)]\right\}]-\ln\left(1-16\epsilon\right)\right),\end{aligned} \tag{54}$$

It is implicit that the arguments of logarithm function above have to be positive–valued as well as the evolutionary distances.

11. Results and discussion

Three types of modifications have been considered for the classical JC and K3ST evolutionary distances: ancestral polymorphisms, purifying selection and *ab initio* removal of deleterious mutations. All three decrease the conventional estimates of the evolutionary distance although the nature of the correction differers. The reduced number of free parameters in JC formulas allows a clear visualization of results.

Figure 2 shows the shape of both uncorrected and corrected (by ancestral polymorphisms and *ab initio* deleterious mutations removal) JC distance (leftmost panels) as well as the explicit value of the corrections (rightmost panels). In the two leftmost panels the curves are the evolutionary distances as function of the mutation fraction. The two rightmost panels illustrate the correction to the JC distance as a function of parameters ϵ or δ. We define the correction to the JC distance as the absolute value of the difference between corrected and uncorrected distances. Panel (a) witnesses that ancestral polymorphisms correction induces a vertical shift that depends only on the ϵ value, a fact that can be seen in panel (b) too. All this can be readily deduced from the formula. Note also that the constraint $12\epsilon < M$ determines the domain of definition of the curves. Bottom panels in Figure 2 refers to *ab initio* removal of deleterious mutations and illustrate how the correction introduced is not a shift. Interestingly, although ϵ and δ represent measurements made at different epochs, similar values yield corrections of same order to JC distance.

Unlike ancestral polymorphisms and *ab initio* deleterious mutations removal, correcting for purifying selection (45) introduces a characteristic time scale τ in the problem. The corrected evolutionary distances may be expressed in units of the characteristic evolutionary scales ν_τ^s, and therefore become dimensionless quantities. Panel (a) of Figure 3 presents corrected $(\bar{\nu}_s/\nu_\tau^s)$ *versus* uncorrected (ν_s/ν_τ^s), JC dimensionless evolutionary distances. The bisecting line (solid) is for $a = 0$ (i.e., no purifying selection), and is given for reference. As far as the value of the parameter a increases the corrected evolutionary distance decreases. For short enough times compared to τ we expect small distance values and we observe that purifying correction acts progressively. For longer times, i.e. large evolutionary distances compared to ν_τ^{jc}, the correction saturates and becomes a constant term. This fact can be more clearly appreciated in panel (b) of Figure 3 where the dimensionless difference $(\nu_s - \bar{\nu}_s)/\nu_\tau^s$ is plotted against ν_s/ν_τ^s. The maximal correction value is determined by the value of a, and it takes place asymptotically in time. This behaviour is consistent with the idea that purifying selection acts only in a limited time span of order τ.

12. Conclusions

We have shown how the matrix formalism for the molecular evolutionary clock leads to the time integration of the substitution model equations in a compact way without preliminary dimensional reductions in phase space. More importantly, the inclusion of further dynamical ingredients has been consistently done for ancestral polymorphisms and purifying selection. The latter has been developed in two different ways. All the three mechanisms lead to smaller evolutionary distances with respect to conventional formulas. As regards ancestral polymorphisms the main results are: equations (32), which is an alternative derivation of a previously published formula [8], as well as the generalizations (34) for JC and (35) for K3ST. The main results concerning purifying selection are equation (45), under the assumption of an exponentially decaying time–dependency in the substitution rates matrix; and (53) and (54) for *ab initio* removal of deleterious mutations. All three equations incorporate the effect of ancestral polymorphisms too.

We have worked out the K3ST model in a phase space of dimension ten and obtained Kimura's original solution after a dimensional reduction. The JC solution is obtained after equating the three substitution rates in K3ST model and carrying out a further dimensional reduction to a phase space of dimension two.

Although JC and K3ST are particularly simple substitution models the full algebraic solution yielded a number of considerations:

1. The full solutions in Tables (2) and (3) witness that K3ST and JC are not time reversible models unless the ancestral base proportions are fixed to the value $1/4$. All the evolutionary distance formulas presented are independent of the stationary time–evolution assumption.
2. There is enough room in the matrix formalism as to incorporate further effects not originally foreseen, preserving essentially the molecular clock scheme. Heterozygosity of the ancestral population may be explicitly accounted for as non–diagonal non–vanishing matrix elements in the initial divergence matrix at divergence time. Including appropriate time–dependence in the rates matrix can simulate the effect of purifying selection and the model can be solved by backward integration. Of course, the methods require the proper estimation of the new parameter values.
3. The effects due to ancestral heterozygosity and *ab initio* deleterious mutations removal lead to corrections to the evolutionary distance that are mathematically different.
4. The correction for purifying selection in (45) is formally the same for JC and K3ST models. The question rises whether this feature holds for higher level substitution models whenever an exponential profile is used to simulate the interference of deleterious mutations.
5. Time varying substitution rates can be handled analytically whenever the time dependency is the same for all nucleotide bases. This is the case of the time variation introduced to deal with purifying selection. If different time dependency for the various mutation types is of interest then the problem becomes much more complex and, as a rule, no analytical solution can be found. Perturbative algebraic methods [19] are needed in those cases to obtain approximate analytical solutions.
6. All these results rely on the assumptions that the molecular clock has been calibrated, and the proportions of ancestral polymorphisms, the fraction of deleterious mutations, as well as the amplitude and time scale of purifying selection have been somehow estimated. To this respect, in [8, 18] proposals to determine ϵ have been developed. As regards purifying selection, estimates for the parameters have also been given [10].

In the past years a huge development of statistical methods for molecular evolution has been carried out. The system of equations that rule the time–evolution of the molecular evolutionary clock is at the core of all of them. We think that there is still room to improve the analytical knowledge of more sophisticated clock models, keeping in mind that every constant rates matrix defines an analytically solvable molecular evolutionary clock and allows to incorporate in it further intricate elements. In this respect it is important to remark that none of the three situations we have considered modify the assumption about the constancy of the substitution rates. Not even the correction for purifying selection introduced via the $J-$shaped function $f(t)$ in Section 7 because it is just an effective way of amending the counting of neutral mutations in the homologous sequences.

The work we have presented has formal and methodological character. We expect the visual evidence we have provided to buttress the potential improvements that the new results can provide. Further assessment will emerge from specific analysis of well chosen nucleotide sequences.

Acknowledgements: GOG thanks The Leverhulme Trust for financial support. JAO work was partially supported by the Spanish MINECO (grant number AYA2016-81065-C2-2).

[1] M. Kimura, The Neutral Theory of Molecular Evolution, Cambridge University Press, Cambridge, 1983.

[2] L. Bromham, D. Penny, The modern molecular clock, Nature Reviews Genetics 4 (2003) 216–224.

[3] J. Felsenstein, Inferring phylogenies, Sinauer Associates, 2004.

[4] Z. Yang, Molecular evolution: a statistical approach, Oxford University Press, Oxford, 2014.

[5] M. Kimura, Estimation of Evolutionary Distances Between Homologous Nucleotide-Sequences, Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences 78 (1981) 454–458.

[6] F. Rodríguez, J. L. Oliver, A. Marín, J. R. Medina, The general stochastic model of nucleotide substitution, Journal of Theoretical Biology 143 (1990) 485–501.

[7] D. Charlesworth, Don't forget the ancestral polymorphisms, Heredity 105 (2010) 509–510.

[8] C. Tuffley, W. T. J. White, M. D. Hendy, D. Penny, Correcting the apparent mutation rate acceleration at shorter time scales under a Jukes–Cantor model, Molecular Biology and Evolution 29 (2012) 3703–3709.

[9] B. D. O'Fallon, A Method to Correct for the Effects of Purifying Selection on Genealogical Inference, Molecular Biology and Evolution 27 (2010) 2406–2416.

[10] P. Soares, L. Ermini, N. Thomson, M. Mormina, T. Rito, A. Röhl, A. Salas, S. Oppenheimer, V. Macaulay, M. B. Richards, Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock, American Journal of Human Genetics. 84 (2009) 740–759.

[11] M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, Journal of Molecular Evolution 16 (1980) 111–120.

[12] T. H. Jukes, C. R. Cantor, Evolution of protein molecules, In: Munro HN (ed) Mammalian protein metabolism, volume 3. Academic Press NY, 1969.

[13] C. Moler, C. Van Loan, Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later, SIAM Review 45 (2003) 3–49.

[14] F. R. Gantmacher, The Theory of Matrices, Vol. 1, AMS Chelsea, Providence, RI, 2000.

[15] S. Tavaré, Some probabilistic and statistical problems in the analysis of DNA sequences, *Lecture on Mathematics in the Life Sciences*, Vol. 17, Springer, Berlin, 1986.

[16] T. Matsumoto, H. Akashi, Z. H. Yang, Evaluation of Ancestral Sequence Reconstruction Methods to Infer Nonstationary Patterns of Nucleotide Substitution, Genetics 200 (2015) 873–890.

[17] M. Kimura, T. Ohta, On the stochastic model for estimation of mutational distance between homologous proteins, Journal of Molecular Evolution 2 (1972) 87–90.

[18] R. Fregel, F. L. Méndez, Y. Bokbot, D. Martín-Socas, M. D. Camalich-Massieu, J. Santana, J. Morales, M. C. Ávila-Arcos, P. A. Underhill, B. Shapiro, G. Wojcik, M. Rasmussen, A. E. R. Soares, J. Kapp, A. Sockell, F. J. Rodríguez-Santos, A. Mikdad, A. Trujillo-Mederos, C. D. Bustamante, Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe, Proceedings of the National Academy of Sciences of the United States of America 115 (2018) 6774–6779.

[19] S. Blanes, F. Casas, J. A. Oteo, J. Ros, The Magnus expansion and some of its applications, Physics Reports 470 (2009) 15–238.

[20] M. Woodhams, Can Deleterious Mutations Explain the Time Dependency of Molecular Rate Estimates?, Molecular Biology and Evolution 23 (2006) 2271–2273.

[21] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, D. E. Knuth, On the Lambert W function, Advances in Computational Mathematics 5 (1996) 329–359.

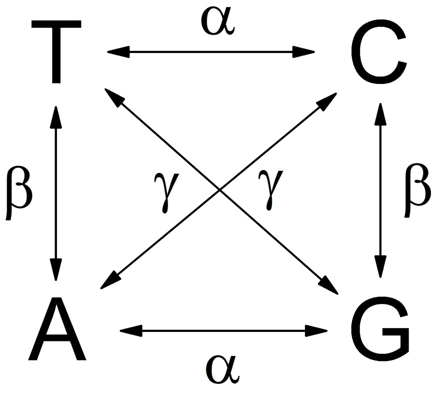

Figure 1: Scheme of K3ST model substitution rates. For JC model, equate all the three rates.

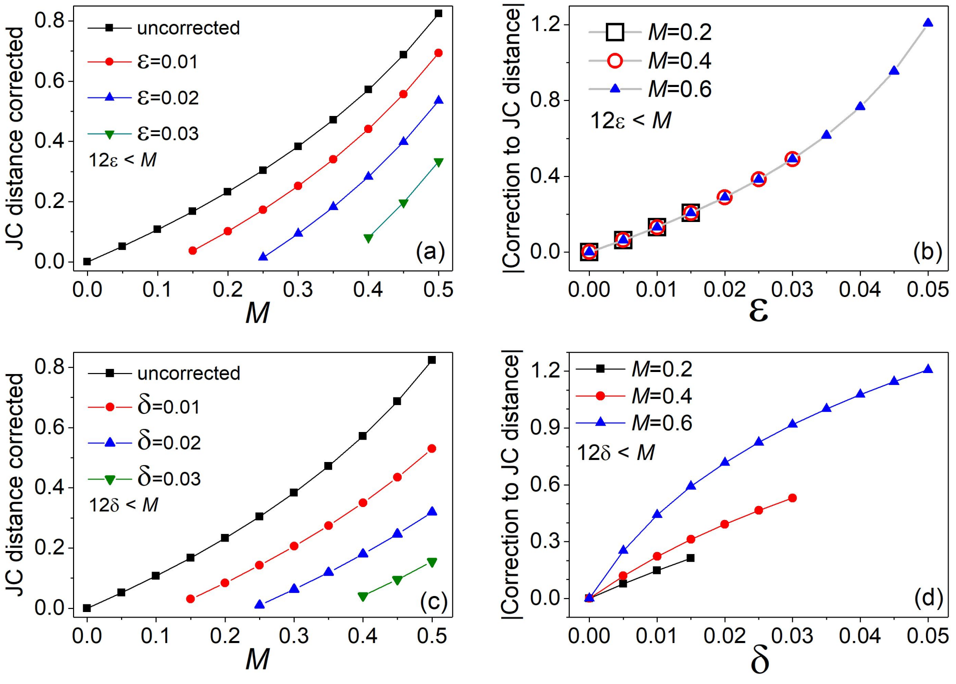

Figure 2: JC evolutionary distance corrected for (a) ancestral polymorphisms (32), and (c) deleterious mutations (49); as a function of the mutation fraction M. Value of the correction to JC distance given by (b) ancestral polymorphisms (32), and (d) *ab initio* deleterious mutations removal (49); as a function of parameters ϵ and δ, respectively. The curves in all four panels are subjected to a constraint that limits their domain of definition.

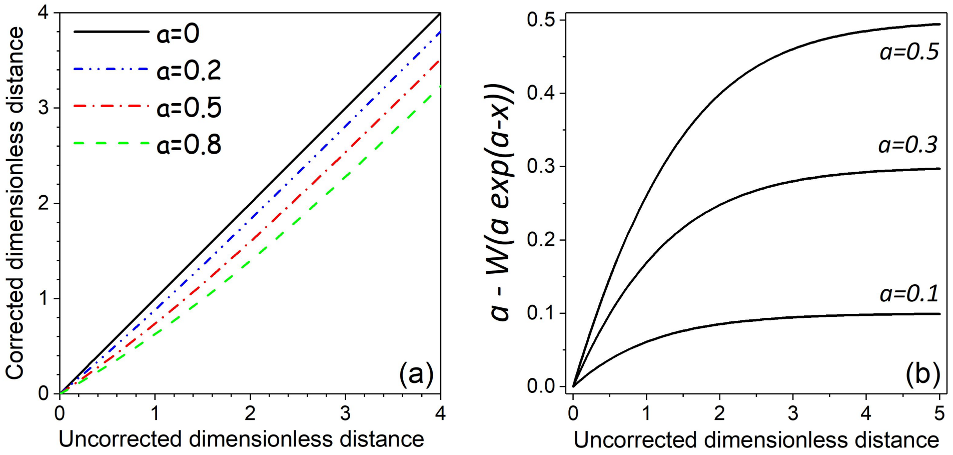

Figure 3: (a) Corrected for purifying selection ($\breve{\nu}_s/\breve{\nu}^s_\tau$) *versus* uncorrected ($\nu_s/\breve{\nu}^s_\tau$) dimensionless evolutionary distances, according to (45). The bisecting line $a = 0$ (i.e., no purifying correction) is given for the sake of reference. Scales are dimensionless. (b) Value of the correction for purifying selection (45): dimensionless value of the difference $(\nu_s - \bar{\nu}_s)/\nu^s_\tau = a - W(a\exp(a - x))$ *versus* uncorrected dimensionless distance ν_s/ν^s_τ, for three values of the amplitude a. In both panels the curves are independent of the index value s in (45), namely the particular evolutionary model and the presence or not of ancestral polymorphisms.

Table 1: Dinucleotide configurations and substitution probability functions.

Base pair:	TT	CC	AA	GG
Probability:	S_1	S_2	S_3	S_4
Base pair:	TC	CT	AG	GA
Probability:	P_1	$\bar{P}_1$	P_2	$\bar{P}_2$
Base pair:	TA	AT	CG	GC
Probability:	Q_1	$\bar{Q}_1$	Q_2	$\bar{Q}_2$
Base pair:	TG	GT	CA	AC
Probability:	R_1	$\bar{R}_1$	R_2	$\bar{R}_2$

Table 2: Time–dependent solutions of the K3ST model divergence matrix. We have defined: $\Gamma = \frac{1}{16}\left(1 + \lambda_{\alpha\beta}^2 + \lambda_{\alpha\gamma}^2 + \lambda_{\beta\gamma}^2\right)$. Rightmost column applies only with initial condition (31), to include ancestral polymorphisms

Base content	Analytical solution	Heterozygosity
TT:	$s_1(t) = \frac{1}{8}\left[-(1+\mathrm{e}^{-2t\alpha})\,\lambda_{\beta\gamma}\sigma_p - (1+\mathrm{e}^{-2t\beta})\,\lambda_{\alpha\gamma}\sigma_q + (1+\mathrm{e}^{-2t\gamma})\,\lambda_{\alpha\beta}\sigma_r\right] + \Gamma$	$-(\lambda_{\alpha\beta} + \lambda_{\alpha\gamma} + \lambda_{\beta\gamma})\epsilon$
CC:	$s_2(t) = \frac{1}{8}\left[-(1+\mathrm{e}^{-2t\alpha})\,\lambda_{\beta\gamma}\sigma_p + (1+\mathrm{e}^{-2t\beta})\,\lambda_{\alpha\gamma}\sigma_q - (1+\mathrm{e}^{-2t\gamma})\,\lambda_{\alpha\beta}\sigma_r\right] + \Gamma$	$-(\lambda_{\alpha\beta} + \lambda_{\alpha\gamma} + \lambda_{\beta\gamma})\epsilon$
AA:	$s_3(t) = \frac{1}{8}\left[(1+\mathrm{e}^{-2t\alpha})\,\lambda_{\beta\gamma}\sigma_p - (1+\mathrm{e}^{-2t\beta})\,\lambda_{\alpha\gamma}\sigma_q - (1+\mathrm{e}^{-2t\gamma})\,\lambda_{\alpha\beta}\sigma_r\right] + \Gamma$	$-(\lambda_{\alpha\beta} + \lambda_{\alpha\gamma} + \lambda_{\beta\gamma})\epsilon$
GG:	$s_4(t) = \frac{1}{8}\left[(1+\mathrm{e}^{-2t\alpha})\,\lambda_{\beta\gamma}\sigma_p + (1+\mathrm{e}^{-2t\beta})\,\lambda_{\alpha\gamma}\sigma_q + (1+\mathrm{e}^{-2t\gamma})\,\lambda_{\alpha\beta}\sigma_r\right] + \Gamma$	$-(\lambda_{\alpha\beta} + \lambda_{\alpha\gamma} + \lambda_{\beta\gamma})\epsilon$
CT+TC:	$p_1(t) = \frac{1}{8}(1 - \lambda_{\alpha\beta}^2 - \lambda_{\alpha\gamma}^2 + \lambda_{\beta\gamma}^2) - \frac{1}{4}(1 - \mathrm{e}^{-2t\alpha})\,\lambda_{\beta\gamma}\sigma_p$	$+2(\lambda_{\alpha\beta} + \lambda_{\alpha\gamma} - \lambda_{\beta\gamma})\epsilon$
AG+GA:	$p_2(t) = \frac{1}{8}(1 - \lambda_{\alpha\beta}^2 - \lambda_{\alpha\gamma}^2 + \lambda_{\beta\gamma}^2) + \frac{1}{4}(1 - \mathrm{e}^{-2t\alpha})\,\lambda_{\beta\gamma}\sigma_p$	$+2(\lambda_{\alpha\beta} + \lambda_{\alpha\gamma} - \lambda_{\beta\gamma})\epsilon$
AT+TA:	$q_1(t) = \frac{1}{8}(1 - \lambda_{\alpha\beta}^2 + \lambda_{\alpha\gamma}^2 - \lambda_{\beta\gamma}^2) - \frac{1}{4}(1 - \mathrm{e}^{-2t\beta})\,\lambda_{\alpha\gamma}\sigma_q$	$+2(\lambda_{\alpha\beta} - \lambda_{\alpha\gamma} + \lambda_{\beta\gamma})\epsilon$
CG+GC:	$q_2(t) = \frac{1}{8}(1 - \lambda_{\alpha\beta}^2 + \lambda_{\alpha\gamma}^2 - \lambda_{\beta\gamma}^2) + \frac{1}{4}(1 - \mathrm{e}^{-2t\beta})\,\lambda_{\alpha\gamma}\sigma_q$	$+2(\lambda_{\alpha\beta} - \lambda_{\alpha\gamma} + \lambda_{\beta\gamma})\epsilon$
GT+TG:	$r_1(t) = \frac{1}{8}(1 + \lambda_{\alpha\beta}^2 - \lambda_{\alpha\gamma}^2 - \lambda_{\beta\gamma}^2) + \frac{1}{4}(1 - \mathrm{e}^{-2t\gamma})\,\lambda_{\alpha\beta}\sigma_r$	$+2(-\lambda_{\alpha\beta} + \lambda_{\alpha\gamma} - \lambda_{\beta\gamma})\epsilon$
AC+CA:	$r_2(t) = \frac{1}{8}(1 + \lambda_{\alpha\beta}^2 - \lambda_{\alpha\gamma}^2 - \lambda_{\beta\gamma}^2) - \frac{1}{4}(1 - \mathrm{e}^{-2t\gamma})\,\lambda_{\alpha\beta}\sigma_r$	$+2(-\lambda_{\alpha\beta} + \lambda_{\alpha\gamma} - \lambda_{\beta\gamma})\epsilon$

Table 3: Time–dependent solutions of the JC model divergence matrix. We have defined: $\gamma = \frac{1}{16}\left[1 + 3\exp(-8\mu t)\right]$. Rightmost column applies only with initial condition (31), to include ancestral polymorphisms.

Base content	Analytical solution	Heterozygosity
TT:	$s_1(t) = \frac{1}{8}(1 + \mathrm{e}^{-2\mu t})\,\mathrm{e}^{-4\mu t}(-\sigma_p - \sigma_q + \sigma_r) + \gamma$	$-3\epsilon\mathrm{e}^{-8\mu t}$
CC:	$s_2(t) = \frac{1}{8}(1 + \mathrm{e}^{-2\mu t})\,\mathrm{e}^{-4\mu t}(-\sigma_p + \sigma_q - \sigma_r) + \gamma$	$-3\epsilon\mathrm{e}^{-8\mu t}$
AA:	$s_3(t) = \frac{1}{8}(1 + \mathrm{e}^{-2\mu t})\,\mathrm{e}^{-4\mu t}(\sigma_p - \sigma_q - \sigma_r) + \gamma$	$-3\epsilon\mathrm{e}^{-8\mu t}$
GG:	$s_4(t) = \frac{1}{8}(1 + \mathrm{e}^{-2\mu t})\,\mathrm{e}^{-4\mu t}(\sigma_p + \sigma_q + \sigma_r) + \gamma$	$-3\epsilon\mathrm{e}^{-8\mu t}$
CT+TC:	$p_1(t) = \frac{1}{8}(1 - \mathrm{e}^{-8\mu t}) - \frac{1}{4}(1 - \mathrm{e}^{-2t\alpha})\,\mathrm{e}^{-4\mu t}\sigma_p$	$+2\epsilon\mathrm{e}^{-8\mu t}$
AG+GA:	$p_2(t) = \frac{1}{8}(1 - \mathrm{e}^{-8\mu t}) + \frac{1}{4}(1 - \mathrm{e}^{-2t\mu})\,\mathrm{e}^{-4\mu t}\sigma_p$	$+2\epsilon\mathrm{e}^{-8\mu t}$
AT+TA:	$q_1(t) = \frac{1}{8}(1 - \mathrm{e}^{-8\mu t}) - \frac{1}{4}(1 - \mathrm{e}^{-2t\mu})\,\mathrm{e}^{-4\mu t}\sigma_q$	$+2\epsilon\mathrm{e}^{-8\mu t}$
CG+GC:	$q_2(t) = \frac{1}{8}(1 - \mathrm{e}^{-8\mu t}) + \frac{1}{4}(1 - \mathrm{e}^{-2t\mu})\,\mathrm{e}^{-4\mu t}\sigma_q$	$+2\epsilon\mathrm{e}^{-8\mu t}$
GT+TG:	$r_1(t) = \frac{1}{8}(1 - \mathrm{e}^{-8\mu t}) + \frac{1}{4}(1 - \mathrm{e}^{-2t\mu})\,\mathrm{e}^{-4\mu t}\sigma_r$	$+2\epsilon\mathrm{e}^{-8\mu t}$
AC+CA:	$r_2(t) = \frac{1}{8}(1 - \mathrm{e}^{-8\mu t}) - \frac{1}{4}(1 - \mathrm{e}^{-2t\mu})\,\mathrm{e}^{-4\mu t}\sigma_r$	$+2\epsilon\mathrm{e}^{-8\mu t}$

Table 4: Solutions of the K3ST divergence matrix evolving in equilibrium regime. The λ's have been defined in equation (7).

Base content	Analytical solution
TT:	$s_1(t) = \frac{1}{16}\left(1 + \lambda_{\alpha\beta}^2 + \lambda_{\alpha\gamma}^2 + \lambda_{\beta\gamma}^2\right)$
CC:	$s_2(t) = \frac{1}{16}\left(1 + \lambda_{\alpha\beta}^2 + \lambda_{\alpha\gamma}^2 + \lambda_{\beta\gamma}^2\right)$
AA:	$s_3(t) = \frac{1}{16}\left(1 + \lambda_{\alpha\beta}^2 + \lambda_{\alpha\gamma}^2 + \lambda_{\beta\gamma}^2\right)$
GG:	$s_4(t) = \frac{1}{16}\left(1 + \lambda_{\alpha\beta}^2 + \lambda_{\alpha\gamma}^2 + \lambda_{\beta\gamma}^2\right)$
CT+TC:	$p_1(t) = \frac{1}{8}\left(1 - \lambda_{\alpha\beta}^2 - \lambda_{\alpha\gamma}^2 + \lambda_{\beta\gamma}^2\right)$
AG+GA:	$p_2(t) = \frac{1}{8}\left(1 - \lambda_{\alpha\beta}^2 - \lambda_{\alpha\gamma}^2 + \lambda_{\beta\gamma}^2\right)$
AT+TA:	$q_1(t) = \frac{1}{8}\left(1 - \lambda_{\alpha\beta}^2 + \lambda_{\alpha\gamma}^2 - \lambda_{\beta\gamma}^2\right)$
CG+GC:	$q_2(t) = \frac{1}{8}\left(1 - \lambda_{\alpha\beta}^2 + \lambda_{\alpha\gamma}^2 - \lambda_{\beta\gamma}^2\right)$
GT+TG:	$r_1(t) = \frac{1}{8}\left(1 + \lambda_{\alpha\beta}^2 - \lambda_{\alpha\gamma}^2 - \lambda_{\beta\gamma}^2\right)$
AC+CA:	$r_2(t) = \frac{1}{8}\left(1 + \lambda_{\alpha\beta}^2 - \lambda_{\alpha\gamma}^2 - \lambda_{\beta\gamma}^2\right)$