A Simple, General Result for the Variance of Substitution Number in Molecular Evolution

Bahram Houchmandzadeh*[,1,2] and Marcel Vallade[2]

[1]CNRS, LIPHY, Grenoble, France

[2]Université Grenoble-Alpes, LIPHY, France

***Corresponding author:** E-mail: bahram.houchmandzadeh@univ-grenoble-alpes.fr.

Associate editor: Claus Wilke

Abstract

The number of substitutions (of nucleotides, amino acids, etc.) that take place during the evolution of a sequence is a stochastic variable of fundamental importance in the field of molecular evolution. Although the *mean* number of substitutions during molecular evolution of a sequence can be estimated for a given substitution model, no simple solution exists for the *variance* of this random variable. We show in this article that the computation of the variance is as simple as that of the mean number of substitutions for both short and long times. Apart from its fundamental importance, this result can be used to investigate the dispersion index R, that is, the ratio of the variance to the mean substitution number, which is of prime importance in the neutral theory of molecular evolution. By investigating large classes of substitution models, we demonstrate that although $R \geq 1$, to obtain R significantly larger than unity necessitates in general additional hypotheses on the structure of the substitution model.

Key words: **substitution number, variance, dispersion index, substitution matrix.**

Article

Introduction

Evolution at the molecular level is the process by which random mutations change the content of some sites of a given sequence (of nucleotides, amino acids, etc.) during time. The number of substitutions n that occur during this time is of prime importance in the field of molecular evolution and its characterization is the first step in deciphering the history of evolution and its many branching. The main observable in molecular evolution, on comparing two sequences, is $\hat{p}$, the fraction of sites at which the two sequences are different. In order to estimate the statistical moments of n, the usual approach is to postulate a substitution model $\mathbf{Q}$ through which $\hat{p}$ can be related to the statistical moments of n. The simplest and most widely used models assume that $\mathbf{Q}$ is site independent, although this constraint can be relaxed (Graur and Li 1999; Yang 2006).

Once a substitution model $\mathbf{Q}$ has been specified, it is straightforward to deduce the *mean* number of substitutions $\langle n \rangle$ and the process is detailed in many textbooks. However, the mean is only the first step in the characterization of a random variable and by itself is a rather poor indicator. The next step in the investigation of a random variable is to obtain its variance V. Surprisingly, no simple expression for V can be found in the literature for arbitrary substitution model $\mathbf{Q}$. The first purpose of this article is to overcome this shortcoming. We show that computing V is as simple as computing $\langle n \rangle$, both for short and long times.

We then apply this fundamental result to the investigation of the dispersion index R, the ratio of the variance to the mean number of substitutions. The neutral theory of molecular evolution introduced by Kimura (1984) supposes that the majority of mutations is neutral (i.e., have no effect on the phenotypic fitness) and therefore substitutions in protein or DNA sequences accumulate at a "constant rate" during evolution, a hypothesis that plays an important role in the foundation of the "molecular clock" (Bromham and Penny 2003; Ho and Duchêne 2014). The original neutral theory postulated that the substitution process is Poissonian, that is, assuming $R = 1$. Since the earliest work on the index of dispersion, it became evident however that R is usually much larger than unity (Gillespie 1989; Ohta 1995; see Cutler [2000] for a review of data). Many alternatives have been suggested to reconcile the "overdispersion" observation with the neutral theory (Cutler 2000). Among these various models, a promising alternative, that of fluctuating neutral space, was suggested by Takahata (1991) which has been extensively studied in various frameworks (Zheng 2001; Bastolla et al. 2002; Wilke 2004; Bloom et al. 2007; Raval 2007).

The fluctuating neutral space model states that the substitution rate m_j^i from state i to state j is a function of both i and j. States i and j can be nucleotides or amino acids, in which case we recover the usual substitution models of molecular evolution discussed above. The states can also be nodes of a neutral graph used to study global protein evolution (Huynen et al. 1996; Bornberg-Bauer and Chan 1999; van Nimwegen et al. 1999). For neutral networks used in the study of protein evolution, Bloom et al. (2007) devised an elegant procedure to estimate the substitution rates. We will show in this article that in general $R \geq 1$ and

the equality is reached only for the most trivial cases. However, producing large R requires additional hypotheses on the structure of substitution rates.

In summary, the problem we investigate in this article is to find a simple and general solution for the variance and dispersion index of any substitution matrix of dimension K. A substitution matrix $\mathbf{Q}$ collects the transition rates m^i_j ($i \neq j$); its diagonal elements $q^i_i = -m^i$ are set such that its columns sum to zero (see below for notations) and designate the rate of leaving state i. Because of this condition, $\mathbf{Q}$ is singular.

Zheng (2001) was the first to use Markov chains to investigate the variance of substitution number as a solution of a set of differential equations. His investigation was further developed by Bloom et al. (2007) who gave the general solution in terms of the spectral decomposition of the substitution matrix; this solution was extended by Raval (2007) for a specific class of matrices used for random walk on neutral graphs. Minin and Suchard (2008) used the same spectral method to derive an analytical form for the generating function of a binary process.

The first step to characterize the substitution number, which as is well known, is to find the equilibrium probabilities π_i of being in a state i, which is obtained by solving the linear system $\sum_i q^i_j \pi_i = 0$ with the additional condition of $\sum_i \pi_i = 1$. Once π_i are obtained, the *mean* substitution number as a function of time is simply $\langle n \rangle = \bar{m}t$ where $\bar{m} = \sum_i m^i \pi_i$ is the weighted average of the "leaving" rates.

We show here that finding the variance necessitates a similar computation. Denoting the weighted deviation of the diagonal elements of $\mathbf{Q}$ from the mean $h_i = (\bar{m} - m^i)\pi_i$, we have to find the solution of the linear system $\sum_i q^i_j r_i = h_j$ with the additional condition $\sum r_i = 0$. For long times, the dispersion index is then simply

$$R = 1 + \frac{2}{\bar{m}} \sum_{i=1}^{K} m^i r_i. \tag{1}$$

For short times, that is, when the mean number of substitutions is small, the result is even simpler:

$$R = 1 + \frac{\nu_m}{\bar{m}^2} \langle n \rangle, \tag{2}$$

where

$$\nu_m = \sum_{i=1}^{K} (\bar{m} - m^i)^2 \pi_i$$

in other words, ν_m is the variance of the diagonal elements of the substitution matrix, weighted by the equilibrium probabilities.

This article is organized as follows: The "Materials and Methods" section contains the theoretical formulation of the problem and its solution that leads to the simple results (1) and (2). The validity of the method is confirmed by Monte Carlo numerical simulations. Generalization to rate heterogeneity is also considered. The "Materials and Methods" section also contains a "Numerical Method" subsection which describes verbally the numerical algorithm provided in supplementary file S1, Supplementary Material online.

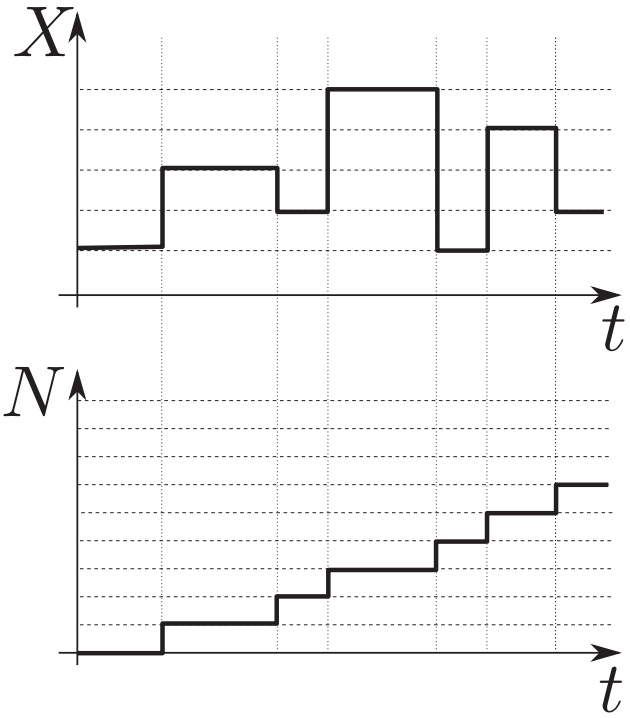

FIG. 1. The random variables X can switch between K states; the counter N of the number of transitions is incremented at each transition of the variable X. The figure above shows one realization of these random variables as a function of time.

In the "Results" section, the simplicity of expressions (1) and (2) are used to study the dispersion index for specific models of nucleotide substitutions widely used in the literature and for general models. We investigate in particular the conditions necessary to produce large R. The last section is devoted to a general discussion of these results and to conclusions. Technical details, such as the proof of $R \geq 1$ are given in the appendices and in supplementary file S2, Supplementary Material online (solution for general initial conditions).

Materials and Methods

Background and Definitions

The problem we investigate in this article is mainly that of counting transitions of a random variable (fig. 1). Consider a random variable X that can occupy K distinct states and let m^i_j ($i \neq j$, $1 \leq i, j \leq K$) be the transition rate from state i to state j. The probability density $p_i(t)$ of being in state i at time t is governed by the Master equation

$$\frac{dp_i}{dt} = -\sum_j m^i_j p_i + \sum_j m^j_i p_j$$

$$= -m^i p_i + \sum_j m^j_i p_j,$$

where $m^i = \sum_j m^i_j$ is the "leaving" rate from state i. We can collect the p_i into a (column) vector $|p\rangle = (p_1, \ldots, p_K)^T$ and write the above equations in matrix notation

$$\frac{d}{dt}|p\rangle = (-\mathbf{D} + \mathbf{M})|p\rangle = \mathbf{Q}|p\rangle, \tag{3}$$

where $\mathbf{D}$ is the diagonal matrix of m^i and $\mathbf{M} = (m^i_j)$ collects the detailed transition rates from state i to state j ($i \neq j$) and has zero on its diagonal. In our notations, the upper (lower) index designates the column (row) of a matrix. The matrix $\mathbf{Q} = -\mathbf{D} + \mathbf{M}$ is called the substitution matrix and its columns sum to zero.

Before proceeding, we explain the notations used in this article. As the matrix $\mathbf{Q}$ is not in general symmetric, a clear distinction must be made between right (column) and left (row) vectors. The Dirac notations are standard and useful for handling this distinction: A column vector $(x_1, \ldots, x_K)^{\mathrm{T}}$ is denoted $|x\rangle$ whereas a row vector $(y^1, \ldots, y^K)$ is denoted $\langle y|$ and $\langle y|x\rangle = \sum_i y^i x_i$ is their scalar product. In some of literature (see Yang 2006), the substitution matrix is the transpose of the matrix used here and the master equation is then written as $\mathrm{d}\langle p|/\mathrm{d}t = \langle p|\mathbf{Q}$ and therefore its rows sum to zero.

By construction, the matrix $\mathbf{Q}$ is singular and has one zero eigenvalue whereas all others are negative. Therefore, as time flows, $|p(t)\rangle >\!\!\rightarrow |\pi\rangle$ where $|\pi\rangle = (\pi_1, \ldots, \pi_K)$ is the equilibrium occupation probability and the zero-eigenvector of the substitution matrix.

$$\mathbf{Q}|\pi\rangle = 0,$$

$$\langle 1|\pi\rangle = 1,$$

where $\langle 1| = (1, \ldots, 1)$, and the second condition expresses that the sum of the probabilities must be 1. Note that by definition, $\langle 1|\mathbf{Q} = 0$, and thus $\langle 1|$ is a zero row eigenvector of the substitution matrix.

Problem Formulation

To count the number of substitutions (fig. 1), we consider the probability densities $p^n_i(t)$ of being in state i after n substitutions at time t. These probabilities are governed by the master equation

$$\frac{\mathrm{d}p^n_i}{\mathrm{d}t} = -m^i p^n_i + \sum_j m^j_i p^{n-1}_j \quad n > 0 \tag{4}$$

$$\frac{\mathrm{d}p^0_i}{\mathrm{d}t} = -m^i p^0_i. \tag{5}$$

We can combine the above equations by setting $p^n_i(t) = 0$ if $n < 0$. Collecting the elements of $(p^n_1, p^n_2, \ldots, p^n_K)^T$ into the vector $|p^n\rangle$, the above equation can then be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}|p^n\rangle = -\mathbf{D}|p^n\rangle + \mathbf{M}|p^{n-1}\rangle. \tag{6}$$

The quantities of interest for the computation of the dispersion index are the mean and the variance of the number of substitutions. The mean number of substitutions at time t is

$$\langle n(t)\rangle = \sum_{i,n} n p^n_i(t).$$

Let us define

$$n_i(t) = \sum_n n p^n_i(t)$$

and collect the partial means n_i into the vector $|n(t)\rangle = (n_1, \ldots, n_K)^{\mathrm{T}}$. The mean is then defined simply as

$$\langle n(t)\rangle = \sum_i n_i(t) = \langle 1|n(t)\rangle. \tag{7}$$

By the same token, the second moment

$$\langle n^2(t)\rangle = \sum_{i,n} n^2 p^n_i(t)$$

can be written in terms of partial second moments $n^2_i = \sum_n n^2 p^n_i(t)$ as

$$\langle n^2(t)\rangle = \langle 1|n^2(t)\rangle, \tag{8}$$

where $|n^2(t)\rangle = (n^2_1, \ldots, n^2_K)^{\mathrm{T}}$. It is straightforward to show (see Appendix A), for the initial condition $|p^n(0)\rangle = |\pi\rangle$, that $|n(t)\rangle$ and $|n^2(t)\rangle$ obey a linear differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}|n\rangle = \mathbf{Q}|n\rangle + \mathbf{D}|\pi\rangle \tag{9}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}|n^2\rangle = \mathbf{Q}|n^2\rangle + 2\mathbf{M}|n\rangle + \mathbf{D}|\pi\rangle. \tag{10}$$

The choice of equilibrium initial condition simplifies the computations and is used through the literature; it is, however, not a necessary requirement. We provide the solution for general initial conditions in supplementary file S2, Supplementary Material online.

Let us define the row vector

$$\langle m| = (m^1, \ldots, m^K) \tag{11}$$

which collects the leaving rates. By definition, $\langle 1|\mathbf{M} = \langle 1|\mathbf{D} = \langle m|$. Multiplying equations (9) and (10) by the row vector $\langle 1|$, and noting that $\langle 1|\mathbf{Q} = \langle 0|$, we get a simple relation for the moments:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle n\rangle = \langle m|\pi\rangle \tag{12}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle n^2\rangle = 2\langle m|n\rangle + \langle m|\pi\rangle. \tag{13}$$

We observe that the mean number of substitutions involves only a trivial integration. Defining the weighted average of the leaving rates as

$$\bar{m} = \langle m|\pi\rangle = \sum_i m^i \pi_i. \tag{14}$$

The mean number of substitution is simply

$$\langle n(t)\rangle = \bar{m}t. \tag{15}$$

To compute the second moment of the substitution number on the other hand, we must solve for $|n\rangle$ using equation (9) and then perform one integration. The next subsection is devoted to the efficient solution of this procedure.

Solution of the Equation for the Moments

One standard way of solving equation (9) would be to express the matrix $\mathbf{Q}$ in its eigenbasis; equation(9) is then

diagonalized and can be formally solved. This is the method used by Bloom et al. (2007) and further refined by Raval (2007) for a specific class of substitution matrices where $m^i_j = 0$ or 1. Bloom et al. (2007, eq. 9) give also a general procedure for finding the variance which can be simplified when an eigenbasis is available.

The problem with the eigenbasis approach is that even when $\mathbf{Q}$ can be diagonalized, this is not the most efficient procedure to find V, as it necessitates the computation of all eigenvalues and row and column eigenvectors of $\mathbf{Q}$ and then the cumbersome summation of their binomial products.

The procedure we follow involves some straightforward, albeit cumbersome linear algebraic operations, but the end result is quite simple. We note that the matrix $\mathbf{Q}$ is singular and has exactly one zero eigenvalue, associated with the row $\langle 1|$ and column $|\pi\rangle$ eigenvectors. The method we use is to isolate the zero eigenvalue by making a round-trip to a new basis. Thus, if we can find a new basis in which the substitution matrix $\mathbf{Q}' = \mathbf{X}^{-1}\mathbf{Q}\mathbf{X}$ takes a lower block triangular form

$$\mathbf{Q}' = \left(\begin{array}{c|c} 0 & 0\cdots 0 \\ \hline \tilde{\alpha} & \tilde{\mathbf{Q}} \end{array}\right) \tag{16}$$

we will have achieved our goal of isolating the zero eigenvalue. The nonsingular matrix $\tilde{\mathbf{Q}}$ is of rank $K-1$ and has the nonzero and negative eigenvalues of $\mathbf{Q}$. As $\langle 1|$ is the known row eigenvalue of $\mathbf{Q}$, we can split the vector space into $\mathcal{B} = \{|u\rangle \,|\, \langle 1|u\rangle = 0\}$ and the space padded by $|\pi\rangle$. It is then straightforward to find the above transfer matrices $\mathbf{X}$ and $\mathbf{X}^{-1}$ for such a transformation:

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & -1 & \cdots & -1 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & & \cdots & 0 & 1 \end{pmatrix};$$

$$\mathbf{X}^{-1} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & & \cdots & 0 & 1 \end{pmatrix}. \tag{17}$$

Under such a transformation, a column vector $|x\rangle = (x_1, x_2, \ldots, x_K)^{\mathsf{T}}$ transforms into

$$|x'\rangle = \mathbf{X}^{-1}|x\rangle = \begin{pmatrix} \sum_i x_i \\ x_2 \\ \vdots \\ x_K \end{pmatrix} = \begin{pmatrix} \sum_i x_i \\ |\tilde{x}\rangle \end{pmatrix}, \tag{18}$$

where the $K-1$ dimensional vector $|\tilde{x}\rangle = (x_2, \ldots, x_K)^{\mathsf{T}}$. In general, we will designate by a $\tilde{}$ all vectors that belong the $K-1$ dimensional space $\mathcal{B}$ in which the linear application $\tilde{\mathbf{Q}}$ operates.

A row vector $\langle y| = (y^1, y^2, \ldots, y^K)$ transforms into

$$\langle y'| = \langle y|\mathbf{X} = (y^1, y^2 - y^1, \ldots, y^K - y^1) = (y^1|\,\langle\tilde{y}|\,),$$

where the $K-1$ dimensional row vector $\langle\tilde{y}| = (y^2 - y^1, \ldots, y^K - y^1)$.

Finally, $\tilde{Q}^i_j = Q^i_j - Q^1_j$ where the elements of $\tilde{\mathbf{Q}}$ have been indexed from 2 to K.

Expressing now the equation (9) for the evolution of first moments in the new basis, we find that

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle n\rangle = \bar{m} \tag{19}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}|\tilde{n}\rangle = \tilde{\mathbf{Q}}|\tilde{n}\rangle + \langle n\rangle|\tilde{\alpha}\rangle + |\tilde{\mu}\rangle, \tag{20}$$

where $|\tilde{n}\rangle = (n_2, \ldots, n_K)^{\mathsf{T}}$, $|\tilde{\mu}\rangle = (m^2\pi_2, \ldots, m^K\pi_K)^{\mathsf{T}}$, and $|\tilde{\alpha}\rangle$ is given in relation (16). Equation (19) is the same as equation (12) and implies that $\langle n\rangle = \bar{m}t$. As $\tilde{\mathbf{Q}}$ is nonsingular (and negative definite), equations (19) and (20) can now readily be solved. Noting that $\mathbf{Q}|\pi\rangle = 0$ implies that $|\tilde{\alpha}\rangle + \tilde{\mathbf{Q}}|\tilde{\pi}\rangle = 0$, the differential equation (20) integrates

$$|\tilde{n}\rangle = \left(\mathbf{I} - e^{\tilde{\mathbf{Q}}t}\right)\tilde{\mathbf{Q}}^{-1}|\tilde{h}\rangle + \langle n\rangle|\tilde{\pi}\rangle, \tag{21}$$

where $|\tilde{h}\rangle = \bar{m}|\tilde{\pi}\rangle - |\tilde{\mu}\rangle$.

To compute the second moment (eq. 13) and the variance, we must integrate the above expression one more time. We finally obtain

$$\begin{aligned}\mathrm{Var}(n) &= \langle n^2\rangle - \langle n\rangle^2 \\ &= \langle n\rangle + 2\left\langle\tilde{m}\right|\left(\mathbf{I}t + \tilde{\mathbf{Q}}^{-1}(\mathbf{I} - e^{\tilde{\mathbf{Q}}t})\right)\tilde{\mathbf{Q}}^{-1}\left|\tilde{h}\right\rangle.\end{aligned} \tag{22}$$

The second term in the right-hand side of the above equation is the excess variance δV with respect to a Poisson process.

Long-Time Behavior

As all eigenvalues of $\tilde{\mathbf{Q}}$ are negative, for large times $\exp(\tilde{\mathbf{Q}}t) \to 0$ and the leading term of the excess variance is therefore

$$\delta V = 2\langle\tilde{m}|\tilde{r}\rangle t, \tag{23}$$

where $|\tilde{r}\rangle$ is the solution of the linear equation

$$\tilde{\mathbf{Q}}|\tilde{r}\rangle = |\tilde{h}\rangle. \tag{24}$$

Returning to the original basis, relation (23) becomes

$$\delta V = 2\langle m|r\rangle t, \tag{25}$$

where $\langle m| = (m^1, m^2, \ldots, m^K)$ is the row vector of the leaving rates and $|r\rangle$ is the solution of the linear equation

$$\mathbf{Q}|r\rangle = |h\rangle, \tag{26}$$

$$\langle 1|r\rangle = 0. \tag{27}$$

$|h\rangle = (h_1, \ldots, h_K)^{\mathrm{T}}$ is the vector of weighted deviation from $\bar{m}$ of the leaving rates m^i:

$$h_i = (\bar{m} - m^i)\pi_i.$$

Finally, for large times, the dispersion index is

$$R = 1 + 2\frac{\langle m|r\rangle}{\bar{m}}, \tag{28}$$

which is the relation (1) given in the introduction.

Figure 2 shows the agreement between the above theoretical results and stochastic numerical simulations.

Two important consequences should be noted. First, it is not difficult to show that $R \geq 1$ for continuous time random processes, which is called the overdispersion of the molecular clock. A demonstration of this theorem for symmetric substitution matrices whose elements are 0 or 1 (adjacency matrices) was given by Raval (2007). We give the general demonstration for general time reversible (GTR) substitution matrices in Appendix B. For arbitrary matrices, extensive numerical results on 10^7 random matrices (fig. 4) have shown $R \geq 1$ with no exception.

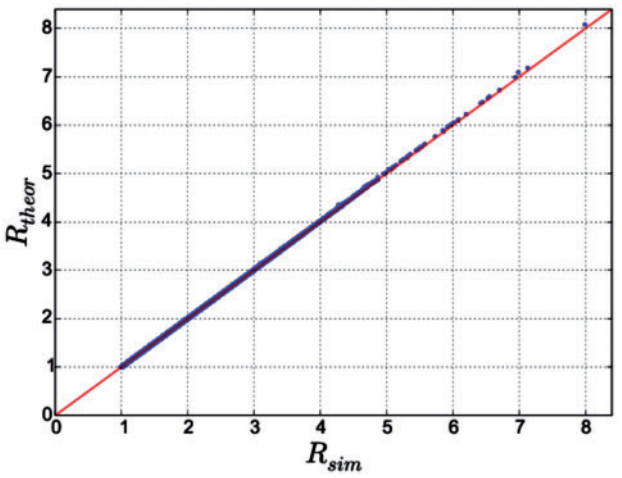

FIG. 2. Comparison between the theoretical result (28) and numerical simulations. 3×10^5 4×4 random (uniform(0, 1)) matrices were generated. For each matrix, a Gillespie algorithm was used to generate 10^6 random paths as a function of time ($t_{\text{final}} = 1{,}000$), from which the dispersion index was computed. In the above figure, each dot corresponds to one random matrix. The mean relative error $(R_{\text{theor}} - R_{\text{sim}})/R_{\text{theor}}$ is 1.3×10^{-3}.

The second consequence of relation (28) is that if all diagonal elements of the substitution matrix are equal (i.e., $m^i = m^j \, \forall i, j$), then the dispersion index is exactly 1 and we recover the property of a normal Poisson process, regardless of the fine structure of $\mathbf{Q}$ and the equilibrium probabilities π_i. This is a sufficient condition. We show that the necessary condition for $R = 1$ is $|h\rangle = |0\rangle$, which, except for the trivial case where some $\pi_i = 0$, again implies the equality of diagonal elements of $\mathbf{Q}$ (see Appendix B).

Note that the expression (28) was obtained for the initial condition $|P(0)\rangle = |\pi\rangle$. As mentioned before, the solution for general initial condition is provided in supplementary file S2, Supplementary Material online, where it is shown that for long times, expression (28) remains valid and does not depend on the choice of initial conditions.

Short-Time Behavior

For short times, that is, when the mean number of substitution is small, we can expand δV given by expression (22) to the second order in time:

$$\delta V = -t^2\langle \tilde{m}|\tilde{h}\rangle + O(t^3) \tag{29}$$

$$= -t^2\sum_{i=2}^{K}(m^i - m^1)(\bar{m} - m^i)\pi_i + O(t^3). \tag{30}$$

Note that the above summation is over $i = 2, \cdots, K$. However, by definition,

$$\sum_{i=1}^{K}(\bar{m} - m^i)\pi_i = 0$$

and hence the sum in relation (30) can be rearranged as

$$\delta V = t^2\sum_{i=1}^{K}(\bar{m} - m^i)^2\pi_i. \tag{31}$$

The sum, which we will denote by ν_m, represents the variance of the diagonal elements of $\mathbf{Q}$, weighted by the equilibrium probabilities. It is more meaningful to express the variance in terms of the mean substitution number. Using relation (15), we therefore have

$$\delta V = \frac{\nu_m}{\bar{m}^2}\langle n\rangle^2. \tag{32}$$

The dispersion index for short times is therefore

$$R = 1 + \frac{\nu_m}{\bar{m}^2}\langle n\rangle.$$

We observe that for short times, the dispersion index increases with the mean number of substitution. This is a generic feature of nontrivial substitution matrices (where diagonal elements are not identical) and has been observed by Ohta (1995).

The dispersion index for all times can also be computed, and an example is given in Appendix C.

Rate Heterogeneity

The substitution models we have considered here can be applied to sequences where the **Q** matrix is identical for all sites. For nucleotide sequences, this can describe the evolution of noncoding sequences or synonymous substitutions. On the other hand, for amino acid substitution, it is well known that some sites are nearly constant or evolve at a slower pace than other sites. A natural extension of the present work would be to take into account substitution matrices that vary among sites, drawing, for example, their scaling factor from a Gamma distribution (Yang 2006). The formalism we have developed in this article can be adapted to such an extension.

Let us define n_S as the substitution number for the sequence:

$$n_S = \sum_{\beta=1}^{L} n_\beta,$$

where L is the sequence length and n_β is the substitution number at site β. If we suppose that sites are not correlated, that is, a transition at one site does not modify the transition rates at other sites, n_β are independent random variables and therefore

$$\langle n_S \rangle = \sum_{\beta=1}^{L} \langle n_\beta \rangle, \tag{33}$$

$$\mathrm{Var}(n_S) = \sum_{\beta=1}^{L} \mathrm{Var}(n_\beta). \tag{34}$$

Consider the simplest case where only the scaling factor varies across sites:

$$\mathbf{Q}_\beta = \lambda_\beta \mathbf{Q},$$

where **Q** is a fixed substitution matrix and λ_β is drawn from a given distribution $f(\lambda)$ with mean $\bar{\lambda}$. The actual form of f is not important for the discussion here. Recalling that (expressions 15 and 25) for large times, $\langle n_\beta \rangle = \lambda_\beta \bar{m} t$ and $\mathrm{Var}(n_\beta) = \lambda_\beta(\bar{m} + 2\langle m|r\rangle)t$, where $\bar{m}$, $\langle m|$, and $|r\rangle$ are defined for the matrix **Q**, expressions (33) and (34) are reduced to

$$\langle n_S \rangle = \bar{m} t \sum_{\beta=1}^{L} \lambda_\beta \approx L\bar{\lambda}\bar{m}t,$$

$$\mathrm{Var}(n_S) = (\bar{m}t + 2\langle m|r\rangle)\sum_{\beta=1}^{L} \lambda_\beta \approx L\bar{\lambda}(\bar{m} + 2\langle m|r\rangle)t,$$

where we have supposed the sequence length L to be large. We observe that the dispersion index in this case is not sensitive to rate variation among sites

$$R_S = \frac{\mathrm{Var}(n_S)}{\langle n_S \rangle} = 1 + 2\frac{\langle m|r\rangle}{\bar{m}},$$

which is the same expression (28) we had before. Therefore, in this simplest case, rate variation among sites cannot increase the dispersion index compared with the homogeneous rate case. For a more general case where all the coefficients of the substitution matrix $\mathbf{Q}_\beta$ are drawn at random from a given distribution, statistical moments of n_S have to be evaluated from expressions (33) and (34) using the exact distribution law $f(q_j^i)$. A similar approach can be used to evaluate the short-time behavior of the dispersion index.

The hypothesis that sites are independent is a simplification and correlations between sites have been used to discover *sectors*, that is, functional domains inside a protein sequence (Halabi et al. 2009). The method presented in this article can in principle be used to study the more general case by considering groups of correlated amino acids as the basic units, which will significantly increase the dimensionality K of the substitution matrix.

Numerical Methods

Numerical simulation of stochastic equations uses the Gillespie (1977) algorithm and is written in C++ language. To compute the dispersion index of a given matrix **Q**, we generate 10^6 random paths over a time period of 1,000. To compare the analytical solutions given in this article (fig. 2) with stochastic simulations (fig. 2), we generated 3.6×10^5 random matrices and numerically computed their dispersion index by the above method. This numerical simulation took approximately 10 days on a 60 core cluster.

All linear algebra numerical computations and all data processing were performed with the high-level Julia language (Bezanson et al. 2014). Computing the analytical dispersion index for the above 3.6×10^5 random matrices took about 5 s on a desktop computer, using only one core.

To generate random GTR matrices, we use the factorization $\mathbf{Q} = \mathbf{S}\Pi^{-1}$ (see Appendix B) which allows for independent generation of the $K(K-1)/2$ elements of the symmetric matrix **S** and $K-1$ elements of Π. For arbitrary matrices, we draw the $K(K-1)$ elements of the matrix. All random generators used in this work are uniform(0,1).

The supplementary file S1, Supplementary Material online, contains the Julia code (15 lines) for computing the dispersion index as given in equation (1). The algorithm can be described as

(1) Find the equilibrium column vector $|\pi\rangle = (\pi_1, \ldots, \pi_K)^T$ by solving the linear system $\mathbf{Q}|\pi\rangle = 0$, with the supplementary condition $\sum_i \pi_i = 1$.
(2) Extract the row vector $\langle m| = (m^1, \ldots, m^K)$ from the diagonal of the **Q** matrix: $m^i = -Q_i^i$.
(3) Compute $\bar{m}$, the rate of variation of the mean substitution number $\bar{m} = \sum_i m^i \pi_i$.
(4) Define the column vector $|h\rangle = (h_1, \ldots, h_K)^T$ whose elements are given by $h_i = (\bar{m} - m^i)\pi_i$.
(5) Find the column vector $|r\rangle = (r_1, \ldots, r_K)^T$ by solving the linear system $\mathbf{Q}|r\rangle = |h\rangle$, with the supplementary condition $\sum_i r_i = 0$.
(6) Compute $\langle m|r\rangle = \sum_i m^i r_i$ and the long-time dispersion index $R = 1 + 2\langle m|r\rangle/\bar{m}$.

Results

Application to Specific Nucleotides Substitution Models

Nucleotide substitution models are widely used in molecular evolution (Graur and Li 1999; Yang 2006), for example, to deduce distances between sequences. Some of these models have few parameters or have particular symmetries. For these models, it is worthwhile to express relation (28) for large times into an even more explicit form and compute the dispersion number as an explicit function of the parameters. We provide below such a computation for some of the most commonly used models.

For the K80 model proposed by Kimura (1980), all diagonal elements of the substitution matrix are equal; hence, relation (28) implies that $R = 1$.

T92 Model

Tamura (1992) introduced a two-parameter model (T92) extending the K80 model to take into account biases in G + C contents. Solving relation (28) explicitly for this model, we find for the dispersion index

$$R = 1 + \frac{2k^2}{k+1}\frac{\theta(1-\theta)(2\theta-1)^2}{1+2k\theta(1-\theta)}. \qquad (35)$$

Here $k = \alpha/\beta$, where α and β are the two parameters of the original T92 model. A similar expression was found by Zheng (2001). For a given k, the maximum value of R is

$$R^* = 1 + \frac{\left(\sqrt{2+k}-\sqrt{2}\right)^2}{k+1}.$$

And it is straightforward to show that in this case

$$R \in [1, 2]$$

although even reaching a maximum value for $R = 1.5$ will necessitate strong asymmetries in the substitution rates (such as $k = 18.8$ and $\theta = 0.063$).

TN93 Model

Tamura and Nei (1993) proposed a generalization of the F81 (Felsenstein 1981) and HKY85 (Hasegawa et al. 1985) models which allows for biases in the equilibrium probabilities, different rates of transition versus transversion and for different rates of transitions. The corresponding substitution matrix is

$$\mathbf{Q}_{\mathrm{TN93}} = \mu \begin{pmatrix} * & k_1\pi_1 & \pi_1 & \pi_1 \\ k_1\pi_2 & * & \pi_2 & \pi_2 \\ \pi_3 & \pi_3 & * & k_2\pi_3 \\ \pi_4 & \pi_4 & k_2\pi_4 & * \end{pmatrix}$$

a specific case of this model where $k_1 = k_2$ corresponds to the HKY85 model, whereas $k_1 = k_2 = 1$ corresponds to that of F81 (also called "equal input"). Solving equation (28) leads to

$$R = 1 + \frac{2}{\bar{m}}\sum_{i<j} C_{ij}(m^i - m^j)^2, \qquad (36)$$

where m^i are the (negative of) diagonal elements of $\mathbf{Q}$, $\bar{m} = \sum_i m^i\pi_i$; C_{ij} are defined as

$$C_{12} = \pi_1\pi_2\frac{1-(k_1-1)(\pi_3+\pi_4)}{1+(k_1-1)(\pi_1+\pi_2)},$$

$$C_{34} = \pi_3\pi_4\frac{1-(k_2-1)(\pi_1+\pi_2)}{1+(k_2-1)(\pi_3+\pi_4)},$$

$$C_{ij} = \pi_i\pi_j \quad \text{for other } i, j.$$

For the specific case $k_1 = k_2 = 1$ (equal input or F81 model), expression (36) takes a particularly simple form

$$R = 1 + \left(\sum_{i<j}\pi_i\pi_j(\pi_i-\pi_j)^2\right) \Big/ \left(\sum_{i<j}\pi_i\pi_j\right) \qquad (37)$$

$$= 1 + 2\frac{\sum_i \pi_i^3 - \left(\sum_i \pi_i^2\right)^2}{1-\sum_i \pi_i^2}. \qquad (38)$$

One can deduce relation (37) from (38) by noting that $\sum_i \pi_i = 1$. As every term of the first sum in relation (37) is smaller than the corresponding term in the second sum:

$$R_{\mathrm{F81}} \in [1, 2].$$

The lower bound is reached for $|\pi\rangle = (1/4)(1,1,1,1)^{\mathrm{T}}$, whereas the upper bound is reached when one of the π_i approaches 1. Zheng (2001) has also computed an expression for the dispersion index for the F81 model; his solution however is rather complicated.

For the general TN93, relation $R \leq 2$ no longer holds. For example, for $|\pi\rangle = (0.6-\epsilon, \epsilon, 0.2, 0.2)$, the dispersion index is

$$R_{\mathrm{TN}} = 0.04 + 0.24k_2 + 6/(6+k_2) + O(\epsilon)$$

and R can become arbitrarily large with appropriate values of k_2.

The simplicity of relation (36) allows for the comprehensive exploration of the hyperplane $\sum_i \pi_i = 1,\ \pi_i > 0$. The results are displayed in figure 3; to obtain large values for R such as $R > 1.5$ necessitates high asymmetries in the transition rates and/or strong biases in equilibrium probabilities of states.

Statistical Investigation of the Dispersion Index and the Influence of Sparseness

The relation (28) can be solved explicitly for general substitution matrices. However, a general substitution matrix of dimension 4 has 11 free parameters (substitution matrices

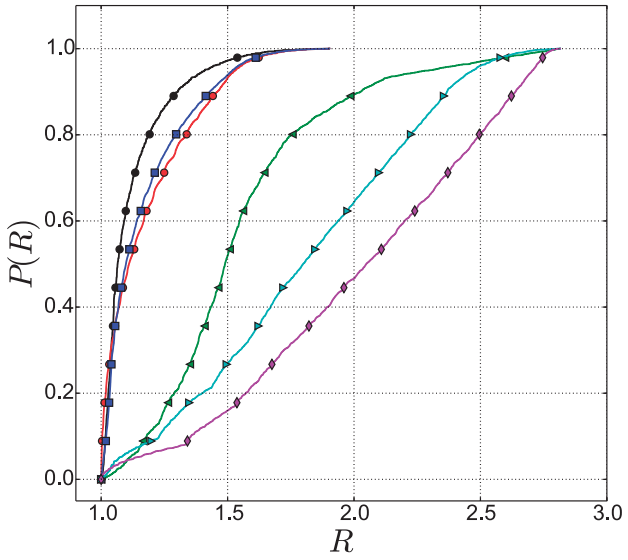

FIG. 3. Cumulative histogram of the dispersion index R of the TN93 model and its specific cases. The three dimensional space $|\pi\rangle = (\pi_1, \pi_2, \pi_3, \pi_4)^T$, $\sum_i \pi_i = 1$ is scanned by steps of $d\pi = 0.025$ ($\approx$11,200 points). For each value of $|\pi\rangle$, the dispersion index of the corresponding substitution matrix $\mathbf{Q}_{TN93}$ is computed from relation (28). Black circles: the F81 model ($k_1 = k_2 = 1$); red diamonds and green left triangles correspond the HKY85 with, respectively, $k_1 = k_2 = 0.1$ and 10; blue squares, cyan right triangles and magenta up triangles correspond to TN93 model with $\{k_1, k_2\} = \{0.1, 1\}, \{1, 10\}, \{0.1, 10\}$, respectively. Permutations of $\{k_1, k_2\}$ lead to the same results and are not displayed. For each substitution matrix, it has been checked that solution (36) and the general solution (28) are identical.

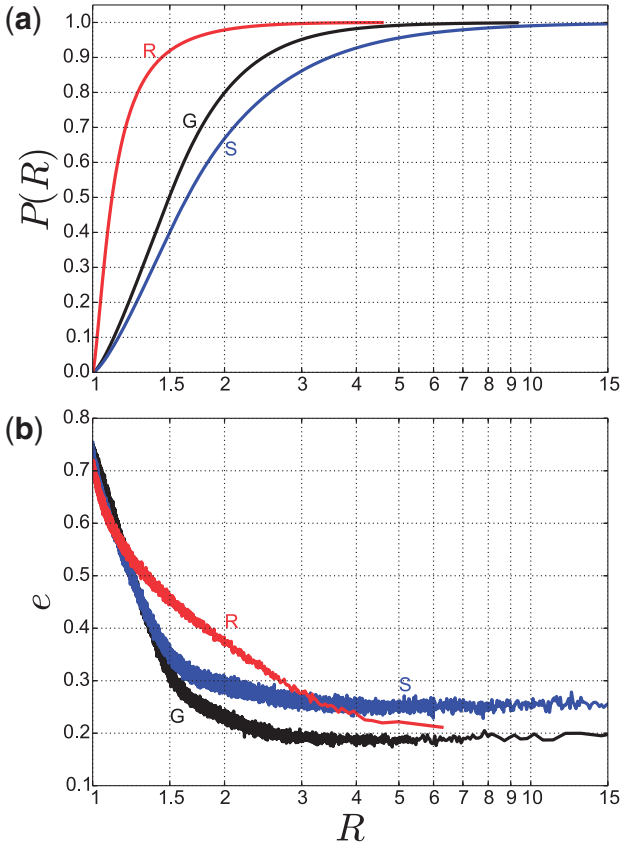

FIG. 4. Statistical study of 4×4 substitution matrices for 1) GTR matrices ("G," black curves), 2) arbitrary matrices ("R," red curves), and 3) sparse GTR matrices ("S," blue curves). In each case, 10^7 matrices are generated and for each matrix, its dispersion index R and its eccentricity $e = \min(\pi_i)$ are computed, where π_i is the equilibrium probability of state i. In each case, the data are sorted by R value to compute the cumulative histogram (a). (b) The relation between e and R. To make visible the statistical relation between e and R, a moving average of size 1,000 data points is applied to the 10^7 sorted (R, e) data in each data set and the result (R_m, e_m) is displayed in the lower plot (b).

are defined up to a scaling parameter); explicit solution of (28) as a function of substitution matrix parameters is rather cumbersome and does not provide insightful information.

An exchangeable (time reversible, GTR) substitution matrix has the additional constraint (Tavaré 1986) $m^i_j \pi_i = m^j_i \pi_j$. GTR matrices are widely used in the literature as they are convenient for computation and are diagonalizable (Yang 2006, Section 2.6).

Considering only exchangeable matrices reduces the number of free parameters to 9, but the parameter space is still too large to be explored systematically.

We can however sample the parameter space by generating a statistically significant ensemble of substitution matrices $\mathbf{Q}$ and get an estimate of the probability distribution of the dispersion index R. The simplicity of relation (28) allows us to generate 10^7 random matrices for each class and compute their associated R in a few minutes with a usual normal computer: Depending on the dimension of $\mathbf{Q}$ (from 4 to 20) this computation takes between 2 and 10 min.

Figure 4*a* shows the cumulative probability $P(R)$ for both arbitrary (R) and GTR (G) matrices, computed from 10^7 matrices in each case. We observe that arbitrary matrices produce statistically low dispersal indices: $P(R > 1.5) = 0.08$ and $P(R > 3) = 2.7 \times 10^{-3}$. The GTR matrices have statistically higher dispersion indices: $P(R > 1.5) = 0.495$ and $P(R > 3) = 0.048$. Still, values larger than $R = 5$, as has been reported in the literature (Cutler 2000), have a very low probability (1.4×10^{-4} for random matrices and 7×10^{-3} for GTR matrices).

We observed in the preceding section that for each class of matrices, high values of R are generally associated with large biases in the equilibrium probabilities, that is, a given state would have a very low equilibrium probability in order to allow for large R. We can investigate how this observation holds for general and GTR matrices. For each $K \times K$ matrix $\mathbf{Q}$ that is generated, we quantify its relative eccentricity by

$$e = K \min_i(\pi_i).$$

The relation between R and e is statistical: Matrices with dispersion index in $[R, R + dR]$ will have a range of e and we display the average e for each small interval (fig. 4*b*). We observe again that high values of the dispersion index in each class of matrices require high bias in equilibrium probabilities of states.

Another effect that can increase the dispersion index of a matrix is its sparseness. This effect was investigated by

Raval (2007) for a random walk on neutral networks, which we generalize here. Until now, we have examined fully connected graphs, that is, substitution processes where the random variable X can jump from any state i to any other state j. For a four-state random variable, each node of the connectivity graph is of degree 3 ($d_G = 3$). This statement may however be too restrictive. Consider, for example, a 4×4 nucleotide substitution matrix for synonymous substitutions. Depending on the identity of the codon to which it belongs, a nucleotide can only mutate to a subset of other nucleotides. For example, for the third codon of tyrosine, only T $\leftrightarrow$ C transitions are allowed, whereas for the third codon of alanine, all substitutions are synonymous. For a given protein sequence, the mean nucleotide synonymous substitution graph is therefore of degree smaller than 3. In general, the degree of each state (node) i is given by the number (minus one) of nonzero elements of the ith column in the associated substitution matrix.

We can investigate the effect of sparseness of substitution matrices on the dispersion index with the formalism developed above. Figure 4 shows the probability distribution of R for GTR matrices with $d_G = 2$. As it can be observed, the dispersion index distribution for GTR matrices is shifted to higher values and $P(R > 5)$ increases 6-fold from 0.007 (for $d_G = 3$) to 0.044 (for $d_G = 2$).

A more insightful model would be 20×20 GTR matrices for amino acid substitutions. We compare the case of fully connected graphs (F) where any amino acid can replace any other one with the case where only amino acids one nucleotide mutation apart can replace each other (nonsynonymous substitution, NS). The average degree of the graph in the latter case is $\bar{d}_G = 7.5$. As before, we generate 10^7 random matrices in each class and compute their statistical properties. We observe again (fig. 5) that the distribution of R is shifted to the right for the NS graphs, where the median is $R_{NS} = 2.48$, compared with $R_F = 1.66$ for fully connected graphs.

For specific amino acid substitution matrices used in the literature such as WAG (Whelan and Goldman 2001), LG (Le and Gascuel 2008) and IDR (Szalkowski and Anisimova 2011), the index of dispersion is 1.253, 1.196 and 1.242, respectively.

Discussion and Conclusion

The substitution process (of nucleotides, amino acids, etc.) and therefore the number of substitutions n that take place during time t are stochastic. One of the most fundamental tasks in molecular evolutionary investigation is to characterize the random variable n from molecular data.

A given model for the substitution process in the form of a substitution matrix $\mathbf{Q}$ enables us to estimate the *mean* number of substitution $\langle n \rangle$ that occurs during a time t. The mean depends only on the diagonal elements of $\mathbf{Q}$ and the equilibrium probabilities of states π_i:

$$\langle n(t) \rangle = -tQ_i^i \pi_i = -t \,\mathrm{tr}(\mathbf{Q}\Pi), \tag{39}$$

where tr() designates the trace operator.

In molecular evolution, the main observable is the probability $p_d(t)$ that two different sequences are different at a given site. Denoting $\mathbf{U}(t) = \exp(t\mathbf{Q})$, and assuming that both sequences are at equilibrium (Yang 2006),

$$p_d(t) = 1 - U_i^i \pi_i = 1 - \mathrm{tr}(\mathbf{U}\Pi). \tag{40}$$

One can estimate $p_d(t)$ from the fraction of observed differences between two sequences $\hat{p}$. By eliminating time in relations (39) and (40), it is then possible to relate the estimators $\hat{d}$ (of $\langle n \rangle$) and $\hat{p}$

$$\hat{d} = f(\hat{p}). \tag{41}$$

For sequences of length L, $\hat{p}$ is given by a binomial distribution $B(L, p)$ and the variance of the distance estimator $\hat{d}$ can be deduced from relation (41). This quantity however is very different from the *intrinsic* variance of the substitution number.

The mean of substitution number, its estimator $\hat{d}$, and the variance of the estimator are only the first step in characterizing a random variable. The next crucial step is to evaluate the variance V of this number. What we have achieved in this article is to find a simple expression for V. In particular, we have shown that for both short and long time scales, the variance V can be easily deduced from $\mathbf{Q}$. For long times, the procedure is similar to deriving the equilibrium probabilities π_i from $\mathbf{Q}$, that is, we only need to solve a linear equation associated with $\mathbf{Q}$ [relation (25)]. For short times, only the diagonal elements of $\mathbf{Q}$ are required to compute V [relation (31)].

A long standing debate in the neutral theory of evolution concerns the value of dispersion index $R = V/\langle n \rangle$. On the one hand, the exact solution of this paper is used to demonstrate that in general, any substitution process given by a

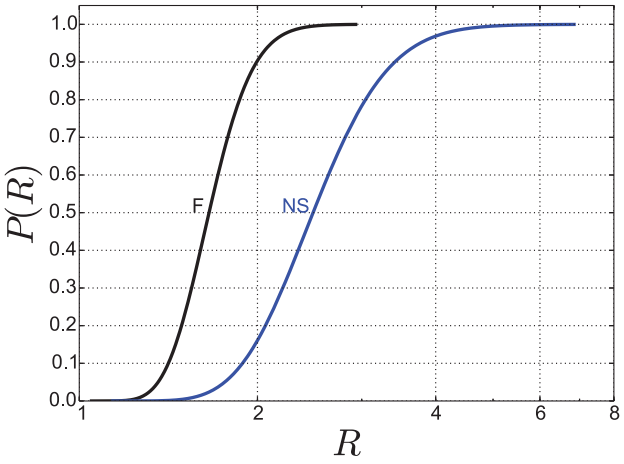

FIG. 5. Cumulative probability of the dispersion index for fully connected (black, F) and nonsynonymous (blue, NS) amino acid 20×20 substitution matrix. For the fully connected matrix, any amino acid can be replaced by any other. For the NS matrices, only amino acids one nucleotide mutation apart can replace each other. For each case, 10^7 20×20 GTR matrices are generated and their dispersion index R is computed. For the NS matrices, transitions are weighted by the number of nucleotide substitutions that can lead from one amino acid to another: There are, for example, 6-nt mutations that transform a phenylalanine into a leucine, but only one mutation that transforms a lysine into an isoleucine.

matrix **Q** is overdispersed, that is, $R \geq 1$, and the equality can be observed only for trivial models where all diagonal elements of **Q** are equal. On the other hand, comprehensive investigation of various substitution models (Results) shows that models that produce R much larger than ≈ 2 generally require strong biases in the equilibrium probabilities of states. One possibility to produce a higher dispersion index is sparse matrices, where the ensemble of possible transitions has been reduced.

Supplementary Material

Supplementary files S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgment

The authors are grateful to Erik Geissler, Olivier Rivoire, and Ivan Junier for fruitful discussions and critical reading of the manuscript.

Appendix A: Mean and Variance Equation

Consider a stochastic system whose transition rates from state **x** to state **y** are $W(x \to y)$. A master equation describes the change in the probability density $P(x,t)$ as a balance between the incoming and outgoing flow:

$$\frac{\mathrm{d}}{\mathrm{d}t}P(\mathbf{x},t) = \sum_{\mathbf{y}} W(\mathbf{y} \to \mathbf{x})P(\mathbf{y},t) - \sum_{\mathbf{y}} W(\mathbf{x} \to \mathbf{y})P(\mathbf{x},t).$$

In the case we study here, the state (n,i) can only be enriched from state $(n-1,j)$ with transition rate m_i^j, as each time a transition occurs, the substitution number is incremented by one unit. On the other hand, the state $(0,i)$ cannot be enriched. The corresponding balance equations are captured in equations (4) and (5).

Obtaining equations of the moments such as (12) and (13) from the Master equation is a standard procedure of stochastic processes (Gardiner 2004; Houchmandzadeh 2009). We give here the outline of the derivation.

Consider the Master equation (6)

$$\frac{\mathrm{d}}{\mathrm{d}t}|p^n\rangle = -\mathbf{D}|p^n\rangle + \mathbf{M}|p^{n-1}\rangle, \tag{42}$$

which is a system of K equations for the $p_i^n(t)$, written in vectorial form. Multiplying each row by n and summing over all n leads, in vectorial form, to

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_n |np^n\rangle = -\mathbf{D}\sum_n |np^n\rangle + \mathbf{M}\sum_n |np^{n-1}\rangle.$$

The term $\sum_n |np^n\rangle$ was defined as the vector of the partial means $|n\rangle$. For the second term, we have

$$\sum_n |np^{n-1}\rangle = \sum_n |(n+1)p^n\rangle = |n\rangle + |p\rangle, \tag{43}$$

where the component p_i of $|p\rangle = \sum_n |p^n\rangle$ is the probability density of the random variable X being in state i, whose dynamics is given by relation (3). For the initial condition $|p(0)\rangle = |\pi\rangle$, we have at all times

$$|p(t)\rangle = |\pi\rangle$$

so the moment equation is

$$\frac{\mathrm{d}}{\mathrm{d}t}|n\rangle = (-\mathbf{D}+\mathbf{M})|n\rangle + \mathbf{M}|\pi\rangle$$

which is relation (9). Note that by definition, $\mathbf{M}|\pi\rangle = \mathbf{D}|\pi\rangle$.

The equation for the second moment (13) is obtained by the same procedure where each row of equation (42) is multiplied by n^2. Higher moments and the probability generating function equations can be obtained by similar computations.

Appendix B: Proof of Overdispersion for GTR Substitution Matrices

As we have seen in relation (28), the dispersion index for long times is

$$R = 1 + 2\frac{\langle m|r\rangle}{\bar{m}}.$$

We must demonstrate that $\langle m|r\rangle \geq 0$ to prove that $R \geq 1$. We give here the proof for GTR matrices. These matrices can be factorized into

$$\mathbf{Q} = \mathbf{S}.\Pi^{-1}, \tag{44}$$

where $\Pi = \mathrm{diag}(\pi_1, \ldots, \pi_k)$ and **S** is a symmetric matrix of positive nondiagonal elements whose columns (and rows) sum to zero. Note that in the literature (Yang 2006), a slightly different factorization is used in the form of $\mathbf{Q} = \Pi.\mathbf{F}$, where **F** is a symmetric matrix (we stress again that in our notation, the substitution matrix is the transpose of that used in most of the literature). The advantage of the factorization (44) is that except for one zero eigenvalue, all other eigenvalues of **S** are negative. **S** can be therefore be written as

$$\mathbf{S} = \sum_{i=2}^{K} \lambda_i |v_i\rangle\langle v_i|, \tag{45}$$

where $|v_i\rangle$ and $\langle v_i|$ are the column and row orthonormal eigenvectors of **S** associated with the eigenvalue λ_i. The pseudoinverse of **S** is defined as

$$\tilde{\mathbf{S}}^{-1} = \sum_{i=2}^{K} \lambda_i^{-1} |v_i\rangle\langle v_i| \tag{46}$$

and it is strictly negative definite.

The vector $|r\rangle$ is the solution of the linear equation

$$\mathbf{Q}|r\rangle = |h\rangle, \tag{47}$$

$$\langle 1|r\rangle = 0, \tag{48}$$

where $h_i = (\bar{m} - m^i)\pi_i$. The general solution of the undetermined equation (47) is therefore

$$|r\rangle = C|\pi\rangle + \Pi\tilde{\mathbf{S}}^{-1}|h\rangle,$$

where the constant C is determined from the condition (48). On the other hand

$$\langle m| = \langle m| - \bar{m}\langle 1| + \bar{m}\langle 1|$$
$$= -\langle h|\Pi^{-1} + \bar{m}\langle 1|.$$

And thus

$$\begin{aligned}\langle m|r\rangle &= -\langle h|\Pi^{-1}|r\rangle + \bar{m}\langle 1|r\rangle \\ &= -\left\langle h|\tilde{\mathbf{S}}^{-1}|h\right\rangle - C\langle h|1\rangle \\ &= -\left\langle h|\tilde{\mathbf{S}}^{-1}|h\right\rangle,\end{aligned} \tag{49}$$

where we have used the fact that $\langle 1|r\rangle = \langle 1|h\rangle = 0$. As $\tilde{\mathbf{S}}^{-1}$ is negative definite,

$$\langle m|r\rangle \geq 0.$$

Moreover, the equality is reached only when $|h\rangle = |0\rangle$, that is, for $\pi_i \neq 0$, only when all diagonal elements of $\mathbf{Q}$ are equal. To see this, we can expand relation (49)

$$\langle m|r\rangle = -\sum_{i=2}^{K} \lambda_i^{-1}\langle v_i|h\rangle^2.$$

The only way to obtain $\langle m|r\rangle = 0$ is to have $\langle v_i|h\rangle = 0$ for $i = 2, \ldots, K$. As on the other hand, $\langle v_1|h\rangle = \langle 1|h\rangle = 0$ we must have $|h\rangle = |0\rangle$.

Appendix C: Dispersion Index for All Times

In "Solution of the Equation for the Moments" section, we gave the long (eq. 28) and short (eq. 32) time solution of the variance. For all the specific models used in the literature, the variance at all times can also be determined explicitly through relation (22)

$$\delta V = 2\left\langle \tilde{m}\right|\left(\mathbf{I}t + \tilde{\mathbf{Q}}^{-1}(\mathbf{I} - e^{\tilde{\mathbf{Q}}t})\right)\tilde{\mathbf{Q}}^{-1}\left|\tilde{h}\right\rangle\rangle. \tag{50}$$

The procedure requires the computation of $\exp(\mathbf{Q}t)$ and is analogous to the determination of $\langle n\rangle$ from sequence dissimilarities (Zheng 2001; Yang 2006).

As an example, consider the equal input model (F81) which we studied in the main text. For this model, the reduced matrix is simply

$$\tilde{\mathbf{Q}} = -\mu\mathbf{I}_3,$$

where $\mathbf{I}_3$ is the 3×3 identity matrix and therefore $\exp(\tilde{\mathbf{Q}}t) = \exp(-\mu t)\mathbf{I}_3$. Relation (50) then becomes

$$\delta V = -\frac{2}{\mu^2}(-1 + \mu t + e^{-\mu t})\langle \tilde{m}|\tilde{h}\rangle.$$

We have previously shown (eq. 31) that generally

$$-\langle \tilde{m}|\tilde{h}\rangle = \sum_{i=1}^{K}(\bar{m} - m^i)^2\pi_i = v_m$$

and for the F81 model,

$$\mu^{-2}v_m = \sum_{i=1}^{K}\pi_i^3 - \left(\sum_{i=1}^{K}\pi_i^2\right)^2.$$

However, the time can be expressed as a function of mean the substitution number. Finally, for the F80 model, and setting $\mu = 1$ without loss of generality, the dispersion index for all times is

$$R(\langle n\rangle) = 1 + 2\left(\frac{\langle n\rangle}{\bar{m}} + e^{-\langle n\rangle/\bar{m}} - 1\right)\frac{v_m}{\langle n\rangle}.$$

References

Bastolla U, Porto M, Roman HE, Vendruscolo M. 2002. Lack of self-averaging in neutral evolution of proteins. *Phys Rev Lett.* 89(20): 208101.

Bezanson J, Edelman A, Karpinski S, Shah VB. 2014. Julia: A Fresh Approach to Numerical Computing. *arXiv*, page 1411.1607.

Bloom JD, Raval A, Wilke CO. 2007. Thermodynamics of neutral protein evolution. *Genetics* 175(1): 255–266.

Bornberg-Bauer E, Chan HS. 1999. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc Natl Acad Sci USA.* 96(19): 10689–10694.

Bromham L, Penny D. 2003. The modern molecular clock. *Nat Rev Genet.* 4(3): 216–224.

Cutler DJ. 2000. The index of dispersion of molecular evolution: slow fluctuations. *Theor Popul Biol.* 57(2): 177–186.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17(6): 368–376.

Gardiner C. 2004. Handbook of stochastic methods: for physics, chemistry and the natural sciences. Berlin (Germany): Springer.

Gillespie DT. 1977. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem.* 81(25): 2340–2361.

Gillespie J. 1989. Lineage effects and the index of dispersion of molecular evolution. *Mol Biol Evol.* 6(6): 636–647.

Graur D, Li WH. 1999. Fundamentals of molecular evolution. Sunderland (MA): Sinauer Associates Inc.

Halabi N, Rivoire O, Leibler S, Ranganathan R. 2009. Protein sectors: evolutionary units of three-dimensional structure. *Cell* 138(4): 774–786.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22(2): 160–174.

Ho SYW, Duchêne S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. *Mol Ecol.* 23(24): 5947–5965.

Houchmandzadeh B. 2009. Theory of neutral clustering for growing populations. *Phys Rev E.* 80(5):051920.

Huynen MA, Stadler PF, Fontana W. 1996. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc Natl Acad Sci USA.* 93(1): 397–401.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol.* 16(2): 111–120.

Kimura M. 1984. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol.* 25(7): 1307–1320.

Minin VN, Suchard MA. 2008. Counting labeled transitions in continuous-time Markov models of evolution. *J Math Biol.* 56(3): 391–412.

Ohta T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J Mol Evol.* 40(1): 56–63.

Raval A. 2007. Molecular clock on a neutral network. *Phys Rev Lett.* 99(13): 138104.

Szalkowski AM, Anisimova M. 2011. Markov models of amino acid substitution to study proteins with intrinsically disordered regions. *PLoS One* 6(5): e20488.

Takahata N. 1991. Statistical models of the overdispersed molecular clock. *Theor Popul Biol.* 39(3): 329–344.

Tamura K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol.* 9(4): 678–687.

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(3): 512–526.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect Math Life Sci.* 17:57.

van Nimwegen E, Crutchfield JP, Huynen M. 1999. Neutral evolution of mutational robustness. *Proc Natl Acad Sci USA*. 96(17): 9716–9720.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*. 18(5): 691–699.

Wilke CO. 2004. Molecular clock in neutral protein evolution. *BMC Genet*. 5(1): 25.

Yang Z. 2006. Computational molecular evolution. Oxford: Oxford Series in Ecology and Evolution.

Zheng Q. 2001. On the dispersion index of a Markovian molecular clock. *Math Biosci*. 172(2): 115–128.