

## RESEARCH ARTICLES

# Quantifying the Stationarity and Time Reversibility of the Nucleotide Substitution Process

Federico Squartini and Peter F. Arndt

Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany

Markov models describing the evolution of the nucleotide substitution process, widely used in phylogeny reconstruction, usually assume the hypotheses of stationarity and time reversibility. Although these models give meaningful results when applied to biological data, it is not clear if the 2 assumptions mentioned above hold and, if not, how much sequence evolution processes deviate from them. To this aim, we introduce 2 sets of indices that can be calculated from the nucleotide distribution and the substitution rates. The stationarity indices (STIs) can be used to test the validity of the equilibrium assumption. The irreversibility indices (IRIs) are derived from the Kolmogorov cycle conditions for time reversibility and quantify the degree of nontime reversibility of a process. We have computed STIs and IRIs for the evolutionary process of 2 lineages, *Drosophila simulans* and *Homo sapiens*. In the latter case, we use a modified form of the indices that takes into account the CpG decay process. In both cases, we find statistically significant deviations from the ideal case of a process that has reached stationarity and is time reversible.

### Introduction

When studying a natural phenomenon, it is a well-established and fruitful practice to disregard some of its properties in order to get a simpler and neater mathematical description. In the first stage, we can use physical and mathematical intuitions to decide what to incorporate and what to eliminate from the description. But once a theory has been laid out, it becomes important to go back to the assumptions previously made and to test in a rigorous way their validity in the phenomenon under study.

In computational evolutionary genomics, one example of this simplification process can be found in the assumptions that are made in the various Markov models of sequence evolution. To be precise, by sequence evolution, we refer to the nucleotide substitution process that leads to the divergence of the DNA sequences of different species originating from a common ancestor. This process can reliably be described using a Markov model (for the basic equations, see the Methods). However, out of historical reasons and computational convenience, several simplifications are usually made inside this framework. The first model, the Jukes–Cantor model or simply JC69, had only one free parameter (Jukes and Cantor 1969). The substitution rate from one nucleotide to any other different nucleotide was assumed to be the same regardless of the particular nucleotides.

A successive model was Kimura's 2-parameter model, also known as K80 (Kimura 1980). This model breaks the complete symmetry present in the JC69, stating that nucleotide evolution has 2 different classes of events. One class is that of "transitions" in which a purine is exchanged with another purine (i.e.,  $A \leftrightarrow G$ ) or a pyrimidine with another pyrimidine (i.e.,  $T \leftrightarrow C$ ). The other class is the one of "transversions" in which a purine is exchanged with a pyrimidine or vice versa (8 possible events:  $A \leftrightarrow T$ ,  $A \leftrightarrow C$ ,  $G \leftrightarrow C$ , and  $G \leftrightarrow T$ ). This reflects biochemical knowledge because the 2 purines, as well as the 2 pyrimidines, have similar chemical structure so that transitions are more likely to happen than transversions.

Key words: Markov models of nucleotide substitutions, stationarity, time reversibility, maximum likelihood estimation, CpG effect.

E-mail: squartin@molgen.mpg.de.

*Mol. Biol. Evol.* 25(12):2525–2535. 2008

doi:10.1093/molbev/msn169

Advance Access publication August 5, 2008

Other models followed that broke more symmetries in the rate matrix: the F81 (Felsenstein 1981), the HKY85 (Hasegawa et al. 1985), the T92 (Tamura 1992), and the TN93 (Tamura and Nei 1993). Eventually, with the formulation of the general time-reversible (GTR) model (Lanave et al. 1984; Tavaré 1986), it was realized that all the models proposed so far (with the exception of the reverse complement symmetric [RCS] model which we will discuss in the next section) were all sharing the symmetry under reversal of time. Later on, several other extensions of these models were introduced, including those which also describe rate heterogeneities along the DNA sequence (Yang 1993; Tuffley and Steel 1998), but they still assume the validity of the GTR model for the evolution of each single nucleotide. For a review, see Liò and Goldman (1998).

It should be noted that models of nucleotide evolution were developed long before genome sequences were available. Researchers had at their disposal the sequences of only small portions of genomes, thus the scarcity of data forced them to use models with as few parameters as possible. In this context, assuming time reversibility and equilibrium in Markov models of nucleotide substitution was an elegant way of restricting the dimensionality of the parameter space. Furthermore, in maximum likelihood calculations, the possibility of rerooting the tree anywhere without affecting the resulting likelihood (the so-called Felsenstein's "pulley principle" [Felsenstein 1981]) leads to an efficient algorithm for calculating the branch lengths of a tree. This speedup is extremely useful when searching the tree space for the maximum likelihood tree.

But is the evolutionary process of nucleotide substitutions really time reversible and in its stationary state? Making such assumptions could cause some important features of genome evolution to be overlooked. If the genome is always in its equilibrium state during evolution, quantities like the average GC content would not evolve in time. However, it was shown by Arndt and Hwa (2005) that, for example, the GC content in the human genome is not in equilibrium and is still evolving. Similar results have also been found for the mouse genome (Duret 2006).

It is the aim of this paper to investigate the stationarity and time-reversal property of the nucleotide substitution

process. We will introduce 2 sets of indices, the stationarity indices (STIs) and the irreversibility indices (IRIs), which can be calculated from the substitution frequencies along one branch in a phylogenetic tree and the nucleotide composition at the node at its end. When nonzero, the indices indicate violations of the basic assumptions mentioned above. We first derive them for Markov models describing the evolution of independent sites. This concept is then extended to models of evolution that also take into account neighbor dependencies along the nucleotide sequence.

It is important to note that although other tests for stationarity and time reversibility have been proposed so far (Saccone et al. 1990; Rzhetsky and Nei 1995; Eyre-Walker 1999; Ababneh et al. 2006) all of them operate on pairs of sequences, which limits their power. For example, situations where a sequence evolved under nonreversible conditions might go undetected as pointed out by Ababneh et al. (2006). Our analysis has the advantage that it tests for stationarity and time reversibility on just any single phylogenetic branch connecting an ancestral node with a more recent one (like, e.g., the branch from the human–chimp common ancestor to present-day human). To compute the indices, the rate matrix has to be estimated using a procedure that does not assume either the time reversibility or the stationarity of the process. To do this, we need at least 3 present-day sequences. A recently developed maximum likelihood analysis can then be employed to reconstruct the substitution frequencies in each branch independently (Duret and Arndt 2008). The relevant details of this method are also exposed in Appendix A.

In the Results, we will calculate the STIs and the IRIs in *Drosophila* and in Primates to check whether the equilibrium and time reversibility properties are granted. For the human genome, we will also take into account the CpG decay process and use an extended version of the IRI.

**Methods**

In the most general Markov model of DNA evolution for a single site, there are 12 distinct substitution processes taking place (Rodríguez et al. 1990). The rates of all these processes,  $\alpha \rightarrow \beta$ , will be denoted  $Q_{\beta\alpha}$ , where greek letters represent the nucleotides: A, C, G, or T. These rates measure the number of substitutions per base pair and per time in a sufficiently small time interval such that multiple substitutions at the same position can be disregarded. For convenience, we write those rates into a  $4 \times 4$  matrix, which we call rate matrix, or generator of the process:

$$Q = \begin{matrix} & \begin{matrix} A & C & G & T \end{matrix} \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{pmatrix} \cdot & Q_{AC} & Q_{AG} & Q_{AT} \\ Q_{CA} & \cdot & Q_{CG} & Q_{CT} \\ Q_{GA} & Q_{GC} & \cdot & Q_{GT} \\ Q_{TA} & Q_{TC} & Q_{TG} & \cdot \end{pmatrix} \end{matrix} \quad (1)$$

The diagonal elements are fixed by the condition that every column adds up to 0, that is,  $Q_{\alpha\alpha} = -\sum_{\beta \neq \alpha} Q_{\beta\alpha}$ . The

time evolution for the probability,  $\rho_\beta$ , to find a nucleotide  $\beta$  is then given by the Master equation:

$$\frac{\partial}{\partial t} \rho_\beta(t) = \sum_\alpha Q_{\beta\alpha} \rho_\alpha(t). \quad (2)$$

The solution to this equation with initial condition  $\rho(0) = (\rho_A(0), \rho_C(0), \rho_G(0), \rho_T(0))^t$  is

$$\rho_\beta(t) = [e^{Qt} \rho(0)]_\beta. \quad (3)$$

The matrix exponential appearing in the preceding equation is called the transition probability matrix of the process for time interval  $t$ :

$$P(t) = e^{Qt}. \quad (4)$$

**The STIs**

The stationary, or equilibrium, state of the process is the probability distribution that does not evolve in time under the evolution defined in equation (2). It is usually denoted as  $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)^t$ , and it can be calculated by solving the following system of linear equations:

$$Q\pi = 0. \quad (5)$$

The easiest way to check whether the process is stationary or not is to define the following indices, which quantify deviations of the present-day nucleotide composition,  $\rho$ , from the equilibrium one,  $\pi$ :

$$\Delta_\alpha = \rho_\alpha - \pi_\alpha. \quad (6)$$

Due to the normalization constraint, only 3 of them are independent. If all of them are equal to zero, that is,  $\Delta_\alpha = 0 \forall \alpha$ , then the process is in its stationary state.

It is important to note that checking for the equality of the nucleotide distribution at different leaf nodes is not a sufficient condition for equilibrium. As an example, all the sequences in the tree could be evolving from a GC-rich state to a GC-poor one with the same rate, in which case they would show the same nucleotide composition even if they are not in equilibrium. Our method does not have such inconveniences and quantifies equilibrium in the most precise way.

We can recast the conditions in a more insightful form if we take independent linear combinations of the  $\Delta_\alpha$  in equation (6) and define:

$$\begin{aligned} STI_1 &= \Delta_C + \Delta_G = \rho_{GC} - \pi_{GC} \\ STI_2 &= \Delta_A - \Delta_T \\ STI_3 &= \Delta_C - \Delta_G, \end{aligned} \quad (7)$$

which we call STIs. The first index is just the difference between the actual GC content,  $\rho_{GC}$ , and the equilibrium GC content,  $\pi_{GC}$ . The second and third equations are reminiscent of the AT skew and GC skew indices. A system is in its stationary state if all STIs vanish.

We further want to quantify whether deviations from 0 of the 3 indices are statistically significant when only

a finite amount of sequence data is available to measure the present-day nucleotide distribution. To achieve this, we compare the distribution of nucleotides,  $\rho_\alpha$ , of a sequence of length  $N$  to the stationary distribution,  $\pi_\alpha$ , using a  $\chi^2$  test with

$$\chi^2 = N \sum_{\alpha} \frac{(\rho_\alpha - \pi_\alpha)^2}{\pi_\alpha}. \quad (8)$$

This quantity follows a  $\chi^2$  distribution with 3 degrees of freedom. Deviations from stationarity are significant (with 95% confidence interval) if  $\chi^2 > 7.8147$ .

### The IRIs

Having analyzed the conditions under which the nucleotide substitution process is in equilibrium, we now turn to the analysis of time reversibility. We first note that a process can be time reversible only if it is already in its equilibrium state (Kelly 1979). From this follows that being out of equilibrium automatically implies a nontime-reversible process. However, we can still ask (and answer) the question whether the process will be time reversible once it will have reached its equilibrium state. To this end, we will introduce the IRIs.

Time reversibility is usually defined through the so-called detailed balance conditions:

$$Q_{\alpha\beta}\pi_\beta = Q_{\beta\alpha}\pi_\alpha \quad \forall \alpha, \beta \in \{A, C, G, T\}. \quad (9)$$

This set of equations must hold if the process is time reversible (Kelly 1979). Once the system has reached the stationary state, the flow of substitutions from nucleotide  $\alpha$  to nucleotide  $\beta$  must be equal to the one from nucleotide  $\beta$  to nucleotide  $\alpha$ . An asymmetry in the 2 flows would define a direction of time and hence break time-reversal symmetry.

Using detailed balance, one needs to check 6 different equations. However, not all of them are independent and therefore we choose instead a different route, using an alternative but equivalent formulation of time reversibility due to Kolmogorov (1936). This leads to a simpler formulation and better highlights the mathematical properties of time reversibility.

A Markov process satisfies the Kolmogorov 3-cycle condition if the product of any 3 matrix elements of the generator satisfies the following equality:

$$Q_{\alpha\gamma}Q_{\gamma\beta}Q_{\beta\alpha} = Q_{\alpha\beta}Q_{\beta\gamma}Q_{\gamma\alpha} \quad \forall \alpha, \beta, \gamma \in \{A, C, G, T\}. \quad (10)$$

The importance of the condition comes from the fact that if it holds and if the process has strictly positive rates, as is the case in the evolutionary process, then the process is time reversible. The proof of this statement is presented in Appendix B. Also notable is the fact that, unlike detailed balance, the Kolmogorov condition does not make use of the equilibrium distribution of the process.

### The Independent Sites Case

In order to check in what case the Markov model defined by equation (1) is also time reversible, we will use Kolmogorov's conditions. We have to consider equalities for the four 3-cycle conditions shown in figure 1a. However, substituting the rate matrix into Kolmogorov conditions, one can immediately check that if any 3 of the 4 conditions are fulfilled, then the fourth holds. That is, there

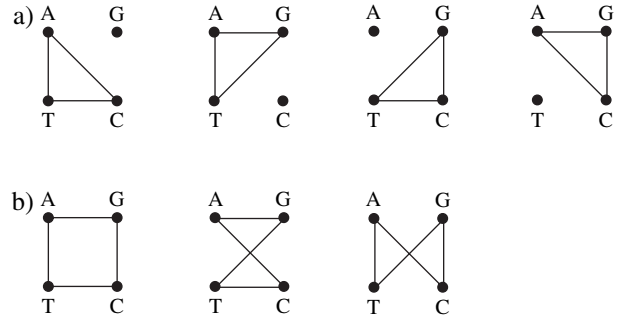


FIG. 1.—All the possible nontrivial 3-cycles (a) and 4-cycles (b) for a Markov model with 4 states.

are only 3 independent 3-cycles, so in order to derive an IRI we could single out 3 of the 4 possible 3-cycles. Instead, we decided to check the equalities on 4-cycles, as there are only 3 nontrivial 4-cycles (see fig. 1b), and they are all independent. This approach is equivalent to the previous one as shown in Appendix B. The process is time reversible if the following conditions:

$$Q_{\alpha\delta}Q_{\delta\gamma}Q_{\gamma\beta}Q_{\beta\alpha} = Q_{\alpha\beta}Q_{\beta\gamma}Q_{\gamma\delta}Q_{\delta\alpha} \quad (11)$$

hold for  $(\alpha, \beta, \gamma, \delta)$  equal to (A, G, C, T), (A, G, T, C), and (A, C, G, T).

It is then straightforward to construct indices out of such equations just by taking the difference of both sides and normalizing it by the sum. We end up with 3 IRIs:

$$\begin{aligned} \text{IRI}_1 &:= \frac{Q_{AG}Q_{GT}Q_{TC}Q_{CA} - Q_{AC}Q_{CT}Q_{TG}Q_{GA}}{Q_{AG}Q_{GT}Q_{TC}Q_{CA} + Q_{AC}Q_{CT}Q_{TG}Q_{GA}} \\ \text{IRI}_2 &:= \frac{Q_{AT}Q_{TG}Q_{GC}Q_{CA} - Q_{AC}Q_{CG}Q_{GT}Q_{TA}}{Q_{AT}Q_{TG}Q_{GC}Q_{CA} + Q_{AC}Q_{CG}Q_{GT}Q_{TA}} \\ \text{IRI}_3 &:= \frac{Q_{AT}Q_{TC}Q_{CG}Q_{GA} - Q_{AG}Q_{GC}Q_{CT}Q_{TA}}{Q_{AT}Q_{TC}Q_{CG}Q_{GA} + Q_{AG}Q_{GC}Q_{CT}Q_{TA}} \end{aligned} \quad (12)$$

The 3 IRIs will thus be comprised in the interval  $[-1, 1]$  and will be simultaneously zero if and only if the system under study evolves time reversibly.

We conclude this section noting that, as we already pointed out in the introduction, evolutionary models traditionally used in the literature belong to a family of nested models that originate from the so-called GTR model, which assumes the following parameterization of the rate matrix:

$$Q_{\text{GTR}} = \begin{matrix} & \begin{matrix} A & C & G & T \end{matrix} \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{pmatrix} \cdot & a\pi_A & b\pi_A & c\pi_A \\ a\pi_C & \cdot & d\pi_C & e\pi_C \\ b\pi_G & d\pi_G & \cdot & f\pi_G \\ c\pi_T & e\pi_T & f\pi_T & \cdot \end{pmatrix} \end{matrix}. \quad (13)$$

The 4  $\pi$ 's appearing in this matrix define the equilibrium distribution of nucleotides; only 3 of them are independent because they are assumed to be normalized. It can easily be checked by substitution of the parameterization (13) in equation (12) that all 3 IRIs vanish for the GTR model,

which therefore is indeed time reversible. The same is true for all its nested submodels, which are mentioned in the Introduction. As expected, the GTR model has 9 free parameters. The 12-dimensional parameter space of the most general model (1) is reduced by 3 dimensions because equating the 3 IRIs to zero yields 3 conditions on the 12 parameters.

### The RCS Case

We now consider a specialized version of the previous model, the RCS model (Wu and Maeda 1987; Lobry 1995; Sueoka 1995). It was originally introduced because it describes very well the evolution of neutrally evolving nucleotide sequences. In this case, the Watson–Crick base coupling between the 2 DNA strands constrains the substitution rates to be symmetric with respect to the transformation which exchange a base with its conjugate (e.g.,  $Q_{AC}$  should be equal to  $Q_{TG}$ ). It follows that the rate matrix has the following form, with 6 parameters:

$$Q_{RCS} = \begin{matrix} & \begin{matrix} A & C & G & T \end{matrix} \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{pmatrix} \cdot & r_{AC} & r_{AG} & r_{AT} \\ r_{GT} & \cdot & r_{CG} & r_{CT} \\ r_{CT} & r_{CG} & \cdot & r_{GT} \\ r_{AT} & r_{AG} & r_{AC} & \cdot \end{pmatrix} \end{matrix} \quad (14)$$

In general, this model is not time reversible. It can be easily checked that the equilibrium distribution has the form  $(1 - \pi_{CG}, \pi_{CG}, \pi_{CG}, 1 - \pi_{CG})/2$  where

$$\pi_{CG} = \frac{r_{GT} + r_{CT}}{r_{AC} + r_{AG} + r_{GT} + r_{CT}}. \quad (15)$$

In this case, the STIs have the following simple form:

$$\begin{aligned} STI_1 &= \rho_{GC} - \pi_{GC} \\ STI_2 &= \rho_A - \rho_T \\ STI_3 &= \rho_C - \rho_G. \end{aligned} \quad (16)$$

It is worth noting that in this case,  $STI_2$  and  $STI_3$  are the unnormalized AT and GC skews. They depend only on the nucleotide composition of the sequence and not on the evolutionary rates. For RCS processes, it can be proven that once these indices or skews vanish, they will stay stationary even if the rate matrix  $Q_{RCS}$  changes in time (Lobry JR and Lobry C 1999). Therefore, the skews can equilibrate even in the presence of RCS rate variations.

To derive an IRI for the RCS model, we substitute the RCS parameterization in equation (12). We find that in this case, we can check time reversibility with just one index:

$$IRI_1 : = \frac{r_{AG}^2 r_{GT}^2 - r_{AC}^2 r_{CT}^2}{r_{AG}^2 r_{GT}^2 + r_{AC}^2 r_{CT}^2}, \quad (17)$$

because  $IRI_2$  and  $IRI_3$  are equal to zero.

### The Case with Neighbor Dependencies

Neighbor dependencies play a significant role in the evolution of vertebrate genomes (Arndt et al. 2003; Hwang and Green 2004). In this case, it is favorable to take into account the CpG decay process as shown in Arndt and Hwa (2005). This is because in presence of methylation, a CpG dinucleotide has an increased mutation rate to a CpA dinucleotide due to the reaction described in Coulondre et al. (1978). The formalism used for describing the sequence evolution must then be appropriately generalized.

The configuration space of a nucleotide sequence of length  $N$  is the Cartesian product of single nucleotide states, having  $4^N$  possible configurations:

$$C = s_1 \times \dots \times s_N, \quad s_i = \{A, C, G, T\}. \quad (18)$$

The nucleotide substitution process in this space will then be described by a  $4^N \times 4^N$  rate matrix, which we assume to have the following form:

$$Q = \sum_{k=1}^N Q_k + \sum_{k=1}^{N-1} Q_{k,k+1}^{CpG}. \quad (19)$$

The first sum represents independent nucleotide evolution and is a sum of tensor products of matrices:

$$Q_k = \underbrace{\mathbb{I} \otimes \dots \otimes \mathbb{I}}_{k-1} \otimes Q \otimes \underbrace{\mathbb{I} \otimes \dots \otimes \mathbb{I}}_{N-k}. \quad (20)$$

Where  $\mathbb{I}$  is the  $4 \times 4$  identity matrix and  $Q$  is given in equation (1). In the rest of this section, we will use the RCS parameterization for  $Q$  (see eq. 14). The second sum in equation (19) represents nearest neighbor dependencies and has the following form:

$$Q_{k,k+1}^{CpG} = \underbrace{\mathbb{I} \otimes \dots \otimes \mathbb{I}}_{k-1} \otimes Q^{CpG} \otimes \underbrace{\mathbb{I} \otimes \dots \otimes \mathbb{I}}_{N-k-1}. \quad (21)$$

$Q^{CpG}$  is a  $16 \times 16$  matrix that model transitions on dinucleotides. In order to include the CpG decay in the model, we parameterize it as follows:

$$Q_{\alpha\beta\alpha\beta}^{CpG} = \begin{cases} r_{CpG} & \text{if } (\alpha' \beta' \alpha\beta) = (CA \ CG) \text{ or } (\alpha' \beta' \alpha\beta) = (TG \ CG) \\ -2r_{CpG} & \text{if } (\alpha' \beta' \alpha\beta) = (CG \ CG) \\ r_{CpG}^{rev} & \text{if } (\alpha' \beta' \alpha\beta) = (CG \ CA) \text{ or } (\alpha' \beta' \alpha\beta) = (CG \ TG), \\ -r_{CpG}^{rev} & \text{if } (\alpha' \beta' \alpha\beta) = (CA \ CA) \text{ or } (\alpha' \beta' \alpha\beta) = (TG \ TG) \\ 0 & \text{otherwise,} \end{cases} \quad (22)$$

where  $r_{CpG}$  is the rate of CpG decay substitutions  $CG \rightarrow CA$  and  $CG \rightarrow TG$  and  $r_{CpG}^{rev}$  is the rate of the corresponding back substitutions. This way we constructed a  $4^N \times 4^N$  rate matrix  $Q$ , whereas the corresponding transition probability matrix  $P = \exp Q$  is computed by matrix exponentiation. In practice, one will only need to handle such matrices for  $N \leq 3$  (for more details, see Appendix A).

To check for the time reversibility of this model of evolution, we should in principle check the Kolmogorov conditions for cycles with vertices in  $C$ . However, the

generator of the dynamics (22) permits only single nucleotide changes at a time and any cycle factorizes and can be decomposed into cycles changing only one site. Therefore, it is sufficient to check Kolmogorov conditions on single nucleotide 3-cycles like we did before, leading to the  $IRI_1$  for the RCS model. In addition to that one has to consider the particular configuration in which a C is followed by a G in the sequence. One example is the 3-cycle  $CG \rightarrow CA \rightarrow CT \rightarrow CG$ . In this case, the factorization is still possible, but it is necessary to add to the total rate the contribution that comes from the CpG deamination process. In summary, there are then 2 IRIs for a process with neighbor dependencies:

$$IRI_1 = \frac{r_{AG}^2 r_{GT}^2 - r_{AC}^2 r_{CT}^2}{r_{AG}^2 r_{GT}^2 + r_{AC}^2 r_{CT}^2}, \quad (23)$$

$$IRI_{CpG} = \frac{r_{GT}^2 (r_{AG} + r_{CpG})^2 - (r_{CT} + r_{CpG}^{rev})^2 r_{AC}^2}{r_{GT}^2 (r_{AG} + r_{CpG})^2 + (r_{CT} + r_{CpG}^{rev})^2 r_{AC}^2}. \quad (24)$$

Note that, as expected, in the absence of neighbor-dependent processes, we have  $IRI_1 = IRI_{CpG}$ .

## Results

In this section, we want to apply our theoretical framework and check whether evolutionary nucleotide substitution processes are in equilibrium and time reversible or not. We will perform this analysis first in flies. In order to measure the STIs and IRIs, it is necessary to estimate the underlying substitution frequencies without assuming time reversibility or stationarity of the evolutionary process. This can be done using a maximum likelihood approach if one has multiple alignments with at least 2 other species available (see Appendix A and Duret and Arndt 2008).

An important feature of the method is that it provides rates for almost all single branches of the phylogenetic tree. This is an advance with respect to the preexisting approaches to the problem that assess stationarity and reversibility for the tree as a whole or as a function of pairs of sequences. The method can further be extended to also include neighbor dependencies, which is important for the analysis of time reversibility in the human lineage discussed in the second part of this section.

### Quantification of Stationarity and Reversibility in Fly

We first measure the STIs and the  $IRI_1$  for the *Drosophila simulans* lineage from the time of the split with *Drosophila sechellia* until the current time, using *Drosophila melanogaster* as the outgroup. Whole-genome alignments of the species are freely available on the Internet (Stark et al. 2007). The genomic sequences have been split into 539 tiles corresponding to 50-kbp long nonoverlapping windows along the *Drosophila* chromosomes. We disregarded all gaps and masked the regions that were annotated as coding sequence in the Ensembl database (Hubbard et al. 2006).

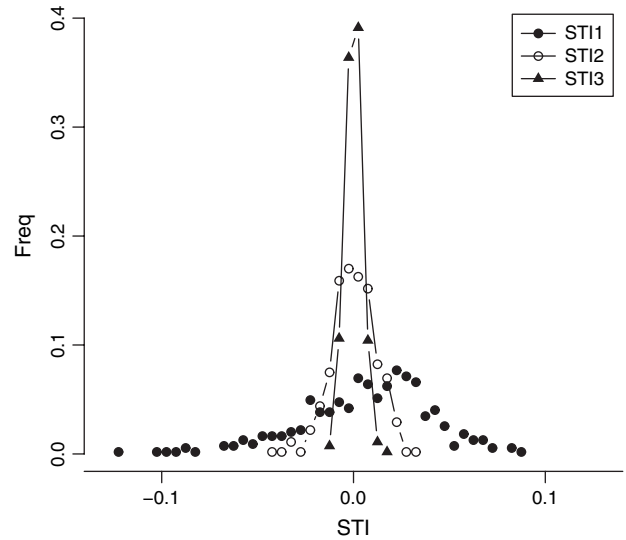


FIG. 2.—The distribution of the  $STI_1$ ,  $STI_2$ , and  $STI_3$  in the *Drosophila simulans* genome. Means and standard deviations are  $STI_1 = 0.007 \pm 0.034$ ,  $STI_2 = 0.000 \pm 0.011$ , and  $STI_3 = 0.000 \pm 0.004$ .

The remaining nucleotides can be regarded to evolve independently from each other and without any significant contribution from the CpG decay process (Arndt and Hwa 2005).

We estimate in each of the 50-kbp windows all 6 free parameters of the RCS model in the *D. simulans* branch. From the inferred substitution rates in each fragment, we have calculated the values of the STIs, thus obtaining the statistical distribution of the indices along the *D. simulans* genome (see fig. 2). The finite variance in the distribution of the indices arises as a statistical effect because we are analyzing finite length sequences in each window and each of them is a realization of a Markov process.

To count how many windows can be assumed to be out of equilibrium, we use the  $\chi^2$  test mentioned in the Methods. Because multiple independent tests are performed, we have applied an appropriate Bonferroni correction, dividing the statistical significance level by the total number of windows. The test accepts the hypothesis of stationarity in only 82 windows and rejects it in 457.

Because the majority of tiles is not in the stationary state, we also analyzed the distribution of the  $IRI_1$  index. The results are summarized in figure 3. We do not present a closed form for the distribution of the  $IRI_1$  in the null case (time reversibility), but the simplicity of the index allowed us to simulate the distribution with little effort. From each window's inferred rate matrix, we constructed an approximated version of the original one with the added property of time reversibility. The construction method uses the fact that any rate matrix  $Q$ , with equilibrium distribution  $\pi$ , can be written in the following way:

$$Q = D(\pi)F = \begin{pmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{pmatrix} \begin{pmatrix} \cdot & F_{12} & F_{13} & F_{14} \\ F_{21} & \cdot & F_{23} & F_{24} \\ F_{31} & F_{32} & \cdot & F_{34} \\ F_{41} & F_{42} & F_{43} & \cdot \end{pmatrix}, \quad (25)$$

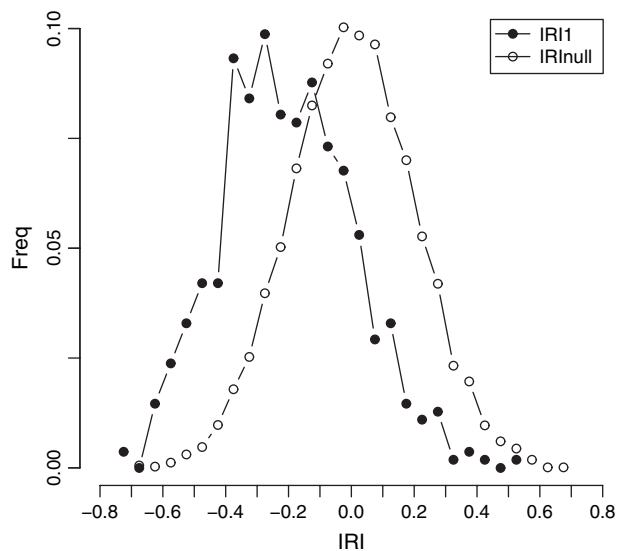


FIG. 3.—The distribution of the IRI in the *Drosophila simulans* genome alongside with the distribution of the IRI for the null model. Means and standard deviations are  $IRI_1 = -0.204 \pm 0.208$  for *D. simulans* and  $IRI_1 = 0.002 \pm 0.197$  for the null model.

for a suitably chosen matrix  $F$ . The dotted elements are again constrained by the fact that the sum of the elements in a column of the rate matrix must be zero,  $\pi_\alpha F_{\alpha\alpha} = -\sum_{\beta \neq \alpha} \pi_\beta F_{\beta\alpha}$ .

We now substitute  $F$  with its symmetrized version and obtain a time-reversible generator  $\hat{Q}$  with the following off-diagonal elements:

$$\hat{Q}_{\alpha\beta} = \pi_\alpha \left[ \frac{F + F^t}{2} \right]_{\alpha\beta}, \quad (26)$$

whereas the diagonal elements are defined as  $\hat{Q}_{\alpha\alpha} = -\sum_{\beta \neq \alpha} \hat{Q}_{\alpha\beta}$ . This generator still has  $\pi$  as equilibrium distribution.

We have used the symmetrized rate matrix to evolve the present-day *D. simulans* sequences contained in each window. We made this in order to simulate evolution under a time-reversible model. We could have used the inferred ancestral *sechellia-simulans* sequence as starting point of the evolution, but because ancestral and present-day sequences have about 13 mismatches per 1,000 bases, this approximation does not affect the following results in any way.

We have then used the RCS model to estimate again the rates, comparing present-day sequences and their evolved counterparts. As a result, we got a second  $IRI_1$  distribution that we have used as null distribution, calling it  $IRI_{Null}$ . The plot is shown in figure 3, and as expected, it is centered in zero.

We performed a 2-sample  $t$ -test to test the null hypothesis that the distributions of  $IRI_1$  and  $IRI_{Null}$  have the same mean. The extremely low  $P$  value of  $10^{-15}$  shows that there is strong evidence against the null hypothesis. In other words, the process is not reversible even when the equilibrium distribution is reached.

## Quantification of Stationarity and Reversibility in the Human Lineage

As a further example, we have measured the STI and IRI for the *Homo sapiens* lineage using a triple alignment of *H. sapiens*, *Pan troglodytes*, and *Macaca mulatta* as an out-group. Whole-genome DNA alignments of these species are available from the Ensembl Web site (Hubbard et al. 2006).

Like in the previous case, we have removed all coding regions and all gaps using Ensembl as a source of annotations. We have split the genome in 2,413 windows of 1-Mbp size. For the analysis of nucleotide substitutions in vertebrates, we have to include neighbor dependencies due to the CpG deamination process and have to use the extended model introduced before.

Distributions of the STIs are shown in figure. A  $\chi^2$  test like the one used above for flies accepts the stationarity hypothesis in only 17 tiles and rejects it in 2,396 tiles. Note that we should in principle also check whether the dinucleotide distribution is stationary. However, because the results show that in the vast majority of tiles already single nucleotides are out of equilibrium, we disregard such an analysis here.

For analyzing time reversibility, it is necessary to use the 2 indices  $IRI_1$  and  $IRI_{CpG}$  introduced in the last part of the Methods. The resulting plots and statistics are shown in figure 5. The same  $t$ -test discussed in the previous section for the equality of the means of  $IRI_1$  and  $IRI_{Null}$  also gives a  $P$  value smaller than  $10^{-15}$ . The  $IRI_{Null}$  distribution in this case has a smaller variance than the  $IRI_1$  distribution. This is because in addition to the variance introduced by finite sequence length, as discussed for *Drosophila*, in the human genome, one finds an intrinsic variation in rates due to its structured nature (Arndt et al. 2005). Time symmetrizing the matrix reduces the dimension of the parameter space and as a consequence reduces heterogeneity in the rates, thus reducing total variance of the  $IRI_1$  in the null model.

## Conclusion

In this paper, we have rigorously analyzed whether the hypotheses of stationarity and time reversibility for the evolution of DNA nucleotide sequences hold. To this end, we have introduced the STIs, which compare the current nucleotide distribution to the stationary one, and the IRIs, which are based on the Kolmogorov cycle conditions for the time reversibility of a Markov process.

We derived explicit expressions of the indices for the general 12-parameter model of nucleotide evolution with independent sites. It is interesting to note that assuming time reversibility, which amounts to setting the IRIs to zero, defines a 9-dimensional submanifold of the 12-dimensional space of all possible models. This manifold is the one spanned by the GTR model and its nested submodels.

As a special case of the previous one, we analyzed RCS models. This particular parameterization arises in a natural way when describing the evolution of neutrally evolving sequences. In this case, it turns out that both STI and IRI have a simpler form. In particular, one needs only one index,  $IRI_1$ , in order to test time reversibility. So, imposing the constraint of time reversibility restricts

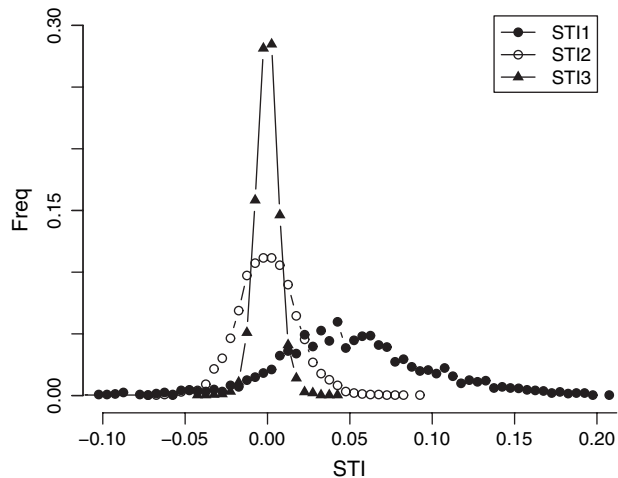


FIG. 4.—The distribution of the  $STI_1$ ,  $STI_2$ , and  $STI_3$  in the human genome. Means and standard deviations are  $STI_1 = 0.052 \pm 0.048$ ,  $STI_2 = 0.000 \pm 0.018$ , and  $STI_3 = 0.000 \pm 0.007$ .

the space of models to a 5-dimensional manifold in the 6-dimensional space of all the possible RCS models. We have successively extended the scope of our study to an evolutionary model that takes into account the CpG decay process, the predominant substitution process in vertebrates.

This approach is complementary to the one using a likelihood ratio test, and it has the advantage that it simultaneously assesses stationarity and time reversibility for all branches of a given phylogeny once the rate matrices have been estimated. On the contrary, a likelihood ratio test requires a comparison of different hypotheses on different branches and a new estimation of the parameters for each of them. When testing for all combinations, the number of likelihood ratio tests required grows exponentially with the number of branches in the phylogeny.

As an application of the theory, we have measured the STI and IRI in 2 different species lineages, *D. simulans* and *H. sapiens*. Using a sliding-window analysis and the maximum likelihood estimation method, we have derived the distributions of STI and  $IRI_1$  for *Drosophila* and of STI,  $IRI_1$ , and  $IRI_{CpG}$  for human. In both cases, we find statistically significant deviations from equilibrium and time reversibility. In *D. simulans*, the values of STI and  $IRI_1$  are close to zero, suggesting that it is legitimate to use a time-reversible Markov model in bioinformatics algorithms, for instance in those used for phylogenetic reconstruction. However, in the human lineage, we find substantial deviations from equilibrium and time reversibility due to the CpG methylation deamination process, in particular  $IRI_{CpG} \approx 1$ . In this case, the lack of equilibrium and time reversibility is an important feature of the probabilistic model and consequently should not be disregarded.

## Appendix A

### Rate Estimation using Maximum Likelihood

To calculate the IRI for nucleotide substitution of a species, we need to estimate the substitution rates  $Q_{\alpha\beta}$  along each branch in a given phylogeny in a manner that does not assume time reversibility or stationarity. This

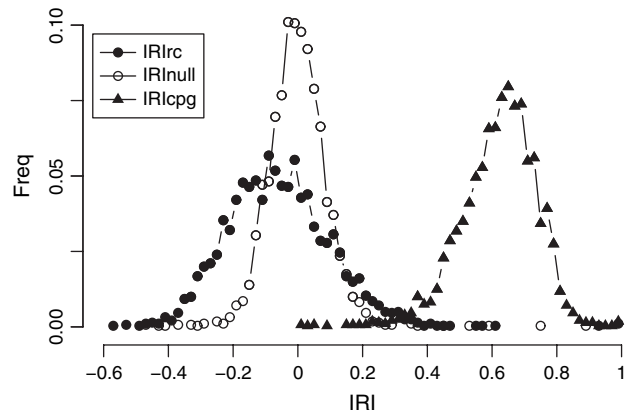


FIG. 5.—The distribution of the  $IRI_1$  and  $IRI_{CpG}$  in the human genome and  $IRI_1$  for the null case. Means and standard deviations are  $IRI_1 = -0.060 \pm 0.161$ ,  $IRI_{CpG} = 0.620 \pm 0.117$ , and  $IRI_1 = -0.002 \pm 0.094$  for the null model.

can be accomplished if we have available DNA sequence alignments with at least 2 other species (Duret and Arndt 2008). Pairwise DNA alignments with a second contemporary (sister) species do not suffice because a base mismatch cannot be unambiguously attributed to 1 of the 2 lineages. This ambiguity does not occur if one uses a triple alignment with an additional third (outgroup) species. In this case, one is able reconstruct the DNA sequence of the last common ancestor of the 2 sister species using maximum likelihood methods. At the same time, this method allows to estimate the substitution frequencies along the branch from the last common ancestor to the sister species. The advantage for our study is that this can be done without making any assumptions on the time reversibility or stationarity of the nucleotide substitution process.

In the following, we first introduce the necessary framework to compute the substitution frequencies from given alignments in cases where nucleotide substitution can be assumed to be neighbor independent, as it is true in flies (Arndt and Hwa 2005). Later, we will extend this method and include neighbor dependencies, which will enable us to study substitutions in vertebrates.

### Neighbor-Independent Substitutions

For a given triple alignment  $\vec{\alpha}^i$  of nucleotide sequences from 3 species,  $i = 1, 2, 3$ , let  $\alpha_k^i$  denote the respective

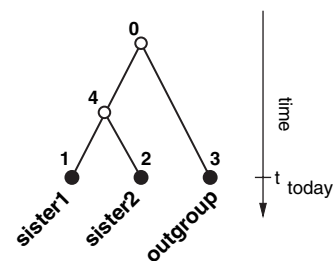


FIG. 6.—The phylogeny of 2 sister species with 1 outgroup species. The numbers next to nodes are used in the text to refer to particular nodes and edges.

homologous nucleotides at position  $k=1, \dots, N$ , where we disregarded all gapped sites. We assume that species number 3 is the outgroup species to the sister species 1 and 2 (see fig. 6). If nucleotide substitution in the 3 species can be assumed to be neighbor independent, then the likelihood of the alignment for given substitution models factorizes and is given by

$$L = \prod_k \sum_{\alpha^0, \alpha^4} \rho_{\alpha^0}^0 [P^{30}]_{\alpha_k^3 \alpha^0} [P^{40}]_{\alpha^4 \alpha^0} [P^{24}]_{\alpha_k^2 \alpha^4} [P^{14}]_{\alpha_k^1 \alpha^4}, \quad (27)$$

where the product runs over all alignment positions and the sums have to be taken over all nucleotides,  $\alpha^0, \alpha^4 \in \{A, C, G, T\}$ . The vector  $\rho^0$  represents the ancestral nucleotide distribution at the root node, and the matrices  $P^{ji}$  are the transition probability matrices for the time evolution along the branches from species  $i$  to species  $j$ . The species number  $i = 0$  refers to the root and  $i = 4$  to the node representing the last common ancestor of the 2 sister species (see fig. 4). These transition probability matrices are parameterized by sets of the substitution frequencies, that is, matrix elements in the corresponding  $Q^{ji}$  matrix, and calculated from them by matrix exponentiation (see eq. 4). We set the length of the time interval  $t$  in each branch to be 1. The inferred substitution frequencies will therefore represent the nucleotide substitution frequencies along the respective branch. This choice of the time unit gives exactly the same results as those obtained with a more standard approach which normalizes that rate matrices such that one branch length unit corresponds to one expected substitution per site. Only the products  $Q^{ji}t$  appear in all expressions and in both cases, the matrix elements  $[P^{ji}]_{\beta\alpha} = [\exp(Q^{ji}t)]_{\beta\alpha}$  give the probabilities for a nucleotide substitution  $\alpha \rightarrow \beta$  along a branch from  $i$  to  $j$ . An equivalent expression for the likelihood was already given by Felsenstein (1981). However, note that the ‘‘pulley principle’’ (Felsenstein 1981) cannot be applied because we consider also irreversible substitution models along the branches of the phylogeny. The computation of the likelihood by ‘‘pruning’’ (Felsenstein 1981) is still possible and allows the effective summation over all configurations of internal nodes.

Please note that we permit different matrices  $Q^{ji}$  along the individual branches in the phylogeny. In the most general case, the likelihood depends on the 4 (=number of branches)  $\times$  12 (=number of parameters along each branch) = 48 substitution frequencies describing the time evolution on each branch and 3 additional parameters for the nucleotide distribution at the root node  $\rho^0$ . In case of a reverse complement substitution model, we need only 6 parameters along each branch (see eq. 14), and the total number of parameters is  $4 \times 6 + 3 = 27$ . For given multiple alignments, it is easy to find the maximum of the likelihood in the 27 dimensional parameter space using the Powell algorithm (Press et al. 1992). Note that the position of the maximum is not unique in the whole parameter space reflecting the fact that the position of the root node cannot be fixed along the 2 branches connected to it (Chang 1996). However, projected onto the subspace parameterizing the substitution models in the 2 sister species, the maximum

is unique and therefore these parameters can be estimated correctly by the maximum likelihood method. Consequently, we use only these parameters in our studies.

### Neighbor-Dependent Substitutions

For the modeling of substitutions in vertebrates, it is necessary to include the neighbor-dependent CpG methylation deamination process that triggers the substitution of cytosine in CpG resulting in TpG or CpA. To have a potentially time-reversible model, we also include the backward processes that mutate CpA or TpG to CpG’s again. Unfortunately, with these neighbor dependencies, the likelihood function does not factorize and is given by

$$L = \sum_{\bar{\alpha}^0, \bar{\alpha}^4} \rho_{\bar{\alpha}^0}^0 [P^{30}]_{\bar{\alpha}^3 \bar{\alpha}^0} [P^{40}]_{\bar{\alpha}^4 \bar{\alpha}^0} [P^{24}]_{\bar{\alpha}^2 \bar{\alpha}^4} [P^{14}]_{\bar{\alpha}^1 \bar{\alpha}^4}. \quad (28)$$

Here  $\rho_{\bar{\alpha}^0}^0$  denotes the probability to have  $\bar{\alpha}^0 \in C$  as the ancestral sequence. The sums in this expression have to be taken over all configurations of sequences at the root  $\bar{\alpha}^0$  and the internal node  $\bar{\alpha}^4$ . The matrices  $P^{ji}$  are  $4^N \times 4^N$  dimensional transition probability matrices describing the evolution of a sequence  $\bar{\alpha}^i$  to  $\bar{\alpha}^j$  along the branch from  $i$  to  $j$ .

In the following, we will assume that the time dynamics is given by neighbor-independent nucleotide substitutions and nearest neighbor-dependent substitutions only. The corresponding generator is given in equation (19). The transition probability matrix is then  $P^{ji} = \exp(tQ^{ji})$ . Without loss of generality, we again set  $t = 1$ .

To maximize the likelihood in equation (28), we introduce a mixed Monte Carlo maximum likelihood (MCML) approach, which combines elements of the 2 methods in a very efficient way: In an iterative fashion, we will first (M-step) estimate the substitution frequencies for a given ancestral sequence at internal nodes (using a maximum likelihood approach) and then (E-step) get a new estimate for the sequence at internal nodes for given the substitution frequencies (using a Monte Carlo approach). This algorithm actually falls into the class of stochastic expectation-maximization (EM) algorithms (McLachlan and Krishnan 1997).

The iteration is initialized setting the sequences at the internal nodes to be the consensus of all its descendant sequences. If nucleotides at one position are not equal in all descendant sequences, one of them is chosen at random. Initializing with a random sequence prolongs but not prevents the convergence of the algorithm to the maximum.

In the M-step, the substitution frequencies (including those for neighbor-dependent processes) are estimated from comparisons of ancestral and daughter sequences as described by Arndt and Hwa (2005). Their method uses a maximum likelihood approach, which accounts for multiple and back substitutions at the same site, and estimates very accurately the substitution frequencies. This approach also admits to include neighbor-dependent processes as the CpG effect and its back mutations.

In the E-step then, we update the ancestral sequences at the internal nodes. To do this, we make use of a Monte

Carlo procedure. We first consider the internal sequence  $\vec{\alpha}^4$ . For each position  $k=1, \dots, N$ , we propose to update the nucleotide  $\alpha_k^4$  by another nucleotide  $\hat{\alpha}_k^4$ . The newly proposed nucleotide is accepted with some probability, which is computed using a local likelihood:

$$L_k^4(\alpha_{k-1}^4 \alpha_k^4 \alpha_{k+1}^4) = P_{\alpha_{k-1}^4 \alpha_k^4 \alpha_{k+1}^4}^{40} \times P_{\alpha_{k-1}^4 \alpha_k^4 \alpha_{k+1}^4}^{14} P_{\alpha_{k-1}^4 \alpha_k^4 \alpha_{k+1}^4}^{24} \quad (29)$$

where the probabilities  $P_{\beta_1 \beta_2 \beta_3 \alpha_1 \alpha_2 \alpha_3}^{ji}$  of substitutions of 3 consecutive nucleotides  $\alpha_1 \alpha_2 \alpha_3$  on node  $i$  to  $\beta_1 \beta_2 \beta_3$  on node  $j$  are given as matrix element of the  $4^3 \times 4^3$  dimensional transition probability matrix  $P^{ji} = \exp Q^{ji}$  describing the time evolution on  $N = 3$  sites with  $Q^{ji}$  given by equation (19). The substitution frequencies along each branch, which fix the corresponding matrices  $Q^{ji}$ , are taken from the estimates in the previous M-step. An update  $\alpha_k^4 \rightarrow \hat{\alpha}_k^4$  is always accepted if the likelihood increases, that is, if the likelihood ratio

$$\lambda = L_k^4(\alpha_{k-1}^4 \hat{\alpha}_k^4 \alpha_{k+1}^4) / L_k^4(\alpha_{k-1}^4 \alpha_k^4 \alpha_{k+1}^4), \quad (30)$$

is larger than one. If this ratio is smaller than one, the substitution is accepted with probability  $\lambda$ . In this case, the (local) likelihood is decreased in order to increase the (global) likelihood in the following M-step.

After the entire internal sequence  $\vec{\alpha}^4$  is updated, the sequence on the root node  $\vec{\alpha}^0$  is updated in a similar fashion. Only the definition of the local likelihood differs and now involves the trinucleotide distribution  $\rho^0(\alpha_1^0 \alpha_2^0 \alpha_3^0)$  of the ancestral sequence  $\vec{\alpha}^0$ :

$$L_k^0(\alpha_{k-1}^0 \alpha_k^0 \alpha_{k+1}^0) = \rho^0(\alpha_{k-1}^0 \alpha_k^0 \alpha_{k+1}^0) \times P_{\alpha_{k-1}^0 \alpha_k^0 \alpha_{k+1}^0}^{40} P_{\alpha_{k-1}^0 \alpha_k^0 \alpha_{k+1}^0}^{30} \quad (31)$$

The trinucleotide distribution is assumed to be homogeneous along the sequence and is estimated from  $\vec{\alpha}^0$  right before starting with the E-step. The transition probabilities are defined as above; the substitution frequencies are given from the estimates in M-step.

This 2 E-and M-step iteration is performed several times until convergence of all the substitution frequencies, and the trinucleotide distribution  $\rho^0(\alpha_1^0 \alpha_2^0 \alpha_3^0)$  is established. In our applications, this happens after about 40 iterations.

By the virtue of the Monte Carlo step, we allow that ancestral sites might not be in their most likely ancestral state. This is done by intention because such situations can actually be observed for sufficiently long sequences. The Monte Carlo step introduces such configurations into the ancestral sequence in as much as they are expected to occur with regard to the substitution model. This is crucial for the accurate estimation of substitution frequencies and ancestral single and dinucleotide frequencies. Note that while the number of those sites that are not in their most likely state is given by the substitution models, their positions are not uniquely defined. Therefore, the ancestral sequence is one representative out of the set of sequences that maximize the likelihood. Although for a general EM algorithm one would require to take the expectation over all possible ancestral sequences (or a sample of those for a Monte Carlo EM algorithm), we rely here on only one represen-

**Table 1**  
**Comparison of Different Models of Nucleotide Substitutions for *Drosophila simulans***

Model	Log $L$	AIC
RCS	-88579.26 (2)	177212.52 (1)
12-parameter	-88570.44 (1)	177242.88 (2)
HKY85	-88612.49 (4)	177246.97 (3)
GTR	-88602.97 (3)	177259.93 (4)
K80	-88647.21 (5)	177316.43 (5)
JC69	-88776.99 (6)	177567.97 (6)

NOTE.—For each model, we report the mean log likelihood, log  $L$ , as well as the mean value of the AIC =  $2p - 2 \log L$ , where  $p$  is the number of parameters of the respective models on the phylogenetic tree. Means are taken over the 539 windows used in the main text. Numbers in brackets report the rank of the corresponding model when sorted by decreasing log  $L$  or increasing AIC.

tative ancestral sequence. This is possible because the average over all positions along the sequence offer an implicit equivalent of the expectation. If only little amounts of sequence data are available, a sampling over different realization of ancestral sequences can easily be incorporated into the MCML approach.

As mentioned above for the neighbor-independent case, the substitution frequencies of edges connected to the root and the trinucleotide distribution of the ancestral sequence  $\vec{\alpha}^0$  cannot be reconstructed. However, the substitution frequencies in the 2 branches for the 2 sister species as well as the nucleotide distribution in the last common ancestor of the 2 sister species are not affected by this ambiguity. For more details and numerical verification of this approach, see Duret and Arndt (2008).

After maximizing the likelihood of a model for given data, the value of the likelihood can also be used to judge whether the use of particular parameterizations is indicated. We performed such a study for the fly data set. A comparison of the RCS model and the GTR model, both of which have 6 free parameters along each branch, came out in favor of the RCS model. Models with more parameters (like the one in eq. 1 with 12 independent parameters) or less parameters (like the HKY85 or JC69 model) compared less favorable with the RCS model when taking into account the total numbers of parameters using the Akaike information criterion (AIC) (see table 1).

## Appendix B

In this appendix, we prove some propositions about the Kolmogorov condition for time reversibility. First of all, we state it for a general Markov process with configuration space  $C$  and transition matrix  $Q$ . Equivalent expressions hold also for the transition probability matrix  $P(t) = \exp Qt$ , which is used in other approaches to evolutionary Markov processes (Barry and Hartigan 1987; Jayaswal et al. 2005).

**Definition 1.** A Markov process is said to satisfy Kolmogorov cycle condition if the following equality on generators holds:

$$Q_{i_1 i_n} Q_{i_n i_{n-1}} \dots Q_{i_2 i_1} = Q_{i_1 i_2} \dots Q_{i_{n-1} i_n} Q_{i_n i_1} \quad \forall i_1, \dots, i_n \in C. \quad (32)$$

The importance of Kolmogorov condition stems from the fact that it can be used to test the reversibility of the process, as proven in the following:

**Proposition 1.** An ergodic Markov process that satisfies Kolmogorov condition is time reversible.

For the proof, see Kelly (1979). Furthermore, the following holds:

**Proposition 2.** If the coefficients of the rate matrix are strictly positive, and if Kolmogorov conditions hold for 3-cycles, then they hold for cycles of arbitrary length.

*Proof.* By induction. It holds trivially for 2 cycles, and it holds by hypothesis for 3-cycles. Then let's show that if it holds for  $n$ -cycles, then it is also valid for  $(n + 1)$  cycles. Let's assume we want to test whether the equality still holds for a chain that has element  $i_{n+1}$  inserted between element  $i_n$  and element  $i_1$ . We multiply both sides by the following factor  $(Q_{i_n i_1} Q_{i_1 i_{n+1}} Q_{i_{n+1} i_n})$ , obtaining

$$\begin{aligned} & (Q_{i_n i_1} Q_{i_1 i_{n+1}} Q_{i_{n+1} i_n}) Q_{i_1 i_n} Q_{i_n i_{n-1}} \cdots Q_{i_2 i_1} \\ & = Q_{i_1 i_2} \cdots Q_{i_{n-1} i_n} Q_{i_n i_1} (Q_{i_n i_1} Q_{i_1 i_{n+1}} Q_{i_{n+1} i_n}), \end{aligned} \quad (33)$$

which after applying Kolmogorov condition for 3-cycles and simplifying leads to

$$Q_{i_1 i_{n+1}} Q_{i_{n+1} i_n} Q_{i_n i_{n-1}} \cdots Q_{i_2 i_1} = Q_{i_1 i_2} \cdots Q_{i_{n-1} i_n} Q_{i_1 i_{n+1}} Q_{i_{n+1} i_n}. \quad (34)$$

So that the equality holds for  $(n + 1)$ -cycles and the proposition is proven.

We now restrict ourselves to a Markov process with only 4 states A, C, G, T and prove the following:

**Proposition 3.** Given a 4 states Markov process with strictly positive rate matrix coefficients, if the conditions:

$$Q_{\alpha\delta} Q_{\delta\gamma} Q_{\gamma\beta} Q_{\beta\alpha} = Q_{\alpha\beta} Q_{\beta\gamma} Q_{\gamma\delta} Q_{\delta\alpha} \quad (35)$$

hold for  $(\alpha, \beta, \gamma, \delta)$  equal to (A, G, C, T), (A, G, T, C), and (A, C, G, T) (see fig. 1b), then Kolmogorov conditions hold for 3-cycles.

*Proof.* It suffices to multiply the generators for the 4-cycles:

$$\begin{aligned} & (Q_{AT} Q_{TC} Q_{CG} Q_{GA}) (Q_{AT} Q_{TG} Q_{GC} Q_{CA}) (Q_{AC} Q_{CT} Q_{TG} Q_{GA}) \\ & = (Q_{AG} Q_{GC} Q_{CT} Q_{TA}) (Q_{AC} Q_{CG} Q_{GT} Q_{TA}) (Q_{AG} Q_{GT} Q_{TC} Q_{CA}). \end{aligned} \quad (36)$$

Simplifying both sides and squaring, we get the equivalence for 1 of the 3-cycles:

$$Q_{GA} Q_{AT} Q_{TG} = Q_{GT} Q_{TA} Q_{AG}. \quad (37)$$

It can be easily seen that exchanging factors between left and right hand side, the remaining 3-cycles can be obtained.

## Literature Cited

Ababneh F, Jermin LS, Ma C, Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics*. 22:1225–1231.

Arndt PF, Hwa T. 2005. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics*. 21:2322–2328.

Arndt PF, Hwa T, Petrov DA. 2005. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *J Mol Evol*. 60:748–763.

Arndt PF, Petrov DA, Hwa T. 2003. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol Biol Evol*. 20:1887–1896.

Barry D, Hartigan JA. 1987. Statistical analysis of hominoid molecular evolution. *Stat Sci*. 2:191–207.

Chang JT. 1996. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Math Biosci*. 134:189–215.

Coullondre C, Miller JH, Farabaugh PJ, Gilbert W. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature*. 274:775–780.

Duret L. 2006. The GC content of primates and rodents genomes is not at equilibrium: a reply to Antezana. *J Mol Evol*. 62:803–806.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*. 4:e1000071.

Eyre-Walker A. 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics*. 152:675–683.

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

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

Hubbard TJP, Aken BL, Beal K, et al. (58 co-authors). 2006. Ensembl 2007. *Nucleic Acids Res*. 35:D610–D617.

Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA*. 101:13994–14001.

Jayaswal V, Jermin LS, Robinson J. 2005. Estimation of phylogeny using a general Markov model. *Evolutionary Bioinformatics*. 1:62–80.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. *Mammalian protein metabolism*. New York: Academic Press. p. 21–123.

Kelly FP. 1979. *Reversibility and stochastic networks*. Chichester (UK): John Wiley & Sons Ltd.

Kimura M. 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *J Mol Evol*. 16:111–120.

Kolmogorov AN. 1936. Zur theorie der markoffschen ketten. *Math Ann*. 112:155–160.

Lanave C, Preparata G, Saccone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. *J Mol Evol*. 20:86–93.

Liò P, Goldman N. 1998. Models of molecular evolution and phylogeny. *Genome Res*. 8:1233–1244.

Lobry JR. 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. *J Mol Evol*. 40:326–330.

Lobry JR, Lobry C. 1999. Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Mol Biol Evol*. 16:719–723.

McLachlan G, Krishnan T. 1997. *The EM algorithm and extensions*. New York: John Wiley & Sons Inc.

Press WH, Flannery BP, Teukolsky SA, Vetterling WT. 1992. *Numerical recipes in C, the art of scientific computing*. Cambridge: Cambridge University Press.

- Rodríguez F, Oliver JL, Marín A, Medina JR. 1990. The general stochastic model of nucleotide substitution. *J Theor Biol.* 142:485–501.
- Rzhetsky A, Nei M. 1995. Tests of applicability of several substitution models for DNA sequence data. *Mol Biol Evol.* 12:131–151.
- Saccone C, Lanave C, Pesole G, Preparata G. 1990. Influence of base composition on quantitative estimates of gene evolution. *Methods Enzymol.* 183:570–583.
- Stark A, Lin MF, Kheradpour P, et al. (43 co-authors). 2007. Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature.* 450:219–232.
- Sueoka N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol.* 40:318–325.
- Tamura K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol.* 9:678–687.
- 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.* 10:512–526.
- Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect Math Life Sci.* 17:57–86.
- Tuffley C, Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. *Math Biosci.* 147:63–91.
- Wu CI, Maeda N. 1987. Inequality in mutation rates of the two strands of DNA. *Nature.* 327:169–170.
- Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol.* 10:1396–1401.

Peter Lockhart, Associate Editor

Accepted July 27, 2008