Eur. Phys. J. E (2012) **35**: 123
DOI 10.1140/epje/i2012-12123-9

THE EUROPEAN PHYSICAL JOURNAL E

Regular Article

Linking the DNA strand asymmetry to the spatio-temporal replication program

II. Accounting for neighbor-dependent substitution rates

A. Baker[1,2], C.L. Chen[3], H. Julienne[1,2], B. Audit[1,2,a], Y. d'Aubenton-Carafa[3], C. Thermes[3], and A. Arneodo[1,2,b]

[1] Université de Lyon, F 69007 Lyon, France
[2] Laboratoire de Physique, ENS Lyon, CNRS, 46 Allée d'Italie, 69364 Lyon Cedex 07, France
[3] Centre de Génétique Moléculaire, CNRS, UPR 3404 associée à l'Université Paris Sud 11, Allée de la Terrasse, 91198 Gif-sur-Yvette, France

Received 8 August 2012 and Received in final form 23 October 2012
Published online: 27 November 2012

Abstract. In paper I, we addressed the impact of the spatio-temporal program on the DNA composition evolution in the case of time homogeneous and neighbor-independent substitution rates. But substitution rates do depend on the flanking nucleotides as exemplified in vertebrates where CpG sites are hypermutable so that the substitution rate $C \to T$ depends dramatically (ten fold) on whether the cytosine belongs to a CG dinucleotide or not. With the specific goal to account for neighbor-dependence, we revisit our minimal modeling of neutral substitution rates in the human genome. When assuming that $r = CpG \to TpG$ and its reverse complement $r^c = CpG \to CpA$ are (by far) the main neighbor-dependent substitution rates, we demonstrate, using perturbative analysis, that neighbor-dependence does not affect the decomposition of the compositional asymmetry into a transcription- and a replication-associated components, the former increases in magnitude with transcription rate and changes sign with gene orientation, whereas the latter is proportional to the replication fork polarity. Indeed the neighbor dependence case differs from the neighbor-independent model by an additional source term related to the CG dinucleotide content in both the transcription and replication-associated components. We finally discuss the case of time-dependent substitution rates confirming as a very general result the fact that the skew can still be decomposed into a transcription- and a replication-associated components.

1 Introduction

Because both these processes require the opening of the DNA double helix and act differently on the two strands, transcription and replication may generate different mutational patterns likely resulting, after long evolutionary time, in a compositional asymmetry of the two DNA strands. As originally observed in prokaryotes, these transcription-associated [1–3] and replication-associated [4–8] compositional asymmetries were further observed in eukaryotes and recently in the human genome with definite evidence of a joined contribution of transcription [9–13] and replication [14–21]. In paper I [22], we have elaborated on a simple but realistic model of time-homogeneous and neighbor-independent neutral substitution rates that describes their dependence upon the replication fork polarity, gene orientation and transcription rate. Then using perturbation theory, we showed that this minimal model theoretically predicts that the compositional asymmetry linearly decomposes into i) a replication-associated component that is proportional to the replication fork polarity and ii) a transcription-associated component that increases in magnitude with transcription rate and changes sign with gene orientation. The validity of this model was demonstrated in the human genome when estimating i) the substitution rates by aligning human and chimpanzee genomes using macaca and orangutan as outgroups and ii) the replication fork polarity as the derivative of genome wide replication timing data for different human cell types. The aim of this paper II is to study to which extent the results reported in paper I still apply when taking into account neighbor dependence since substitution rates are well known to depend on flanking nucleotides [23,24]. Substitutional patterns have also been demonstrated to depend on time [23, 25]. In this case all substitution rates and all the parameters derived from them will depend explicitly on time making the mathematical treatment rather difficult to handle. Notably, in contrast to the time-homogeneous case, the nucleotide composition no longer necessarily converges

[a] e-mail: benjamin.audit@ens-lyon.fr
[b] e-mail: alain.arneodo@ens-lyon.fr

towards its equilibrium value. As a first step towards future theoretical developments, we will provide hints and indications how to formally take into account the time dependence.

The manuscript is organized as follows. Section 2 is devoted to the general formalism for DNA composition evolution when taking into account neighbor-dependent substitution rates. We mainly follow the model introduced by Arndt *et al.* [26] and review some of its general properties. In sect. 3, we focus our study on the $CpG \to TpG$ substitution only, which is by far (13 fold) more frequent than $C \to T$, the most frequent single nucleotide substitution [27,28]. We show that the $(CpG \to TpG)^a$ asymmetry decomposes into transcription- and replication-associated components consistently with the minimal model presented in paper I. Using perturbation theory, we demonstrate and confirm that the compositional asymmetry in the human genome is compliant with the model proposed in paper I: the replication-associated asymmetry is proportional to the replication fork polarity whereas the transcription-associated adds to the previous one and changes sign with gene orientation. In sect. 4, we address the issue of time dependence of substitution rates. We conclude in sect. 5 by discussing the implications of this theoretical study on previous and future works.

2 DNA composition evolution

2.1 Neighbor-dependent model

According to [26], the DNA sequence mainly evolves by two processes: i) single-nucleotide (neighbor-independent) substitution rates and ii) dinucleotide (neighbor-dependent) substitution rates. Along that line, let us define a single-nucleotide substitution rate matrix M and a dinucleotide substitution rate matrix Q. We keep the same definition for M as in paper I: for nucleotides $i, a \in \{T, A, G, C\}$, the element M_{ia} is the (neighbor-independent) substitution rate $a \to i$. For $i, j \in \{T, A, G, C\}$ and $a, b \in \{T, A, G, C\}$, the element $Q_{ij,ab}$ is the (neighbor-dependent) substitution rate $ab \to ij$. As for any transition rate matrix, the sums over rows are null for both M and Q

$$\sum_i M_{ia} = 0 \qquad \text{and} \qquad \sum_{ij} Q_{ij,ab} = 0. \tag{1}$$

The evolution of the composition is then given by

$$\frac{\mathrm{d}}{\mathrm{d}t} X_i(t) = \sum_a M_{ia} X_a(t) + \sum_{abc} Q_{ia,bc} X_{bc}(t) + \sum_{abc} Q_{ai,bc} X_{bc}(t), \tag{2}$$

where $X_i(t)$ is still the frequency (or probability) of nucleotide i at time t, and $X_{ij}(t)$ the frequency (or probability) of dinucleotide ij at time t. The first term accounts for the single nucleotide substitutions. The second and third terms account for all the dinucleotide substitutions that give rise to the nucleotide i. A dinucleotide bc can substitute into a dinucleotide ia (second term) with a nucleotide i on the first base, or into a dinucleotide ai (third term) with a nucleotide i on the second base.

2.1.1 The neighbor-dependent model is not a closed system

According to eq. (2), the time evolution of the composition in nucleotides $X_i(t)$ depends on the composition in dinucleotides $X_{ij}(t)$. Hence, in order to solve this equation, we need to determine the composition in dinucleotides. The evolution of the composition in dinucleotides is given by

$$\frac{\mathrm{d}}{\mathrm{d}t} X_{ij}(t) = \sum_{ab} [M \otimes \mathbb{I} + \mathbb{I} \otimes M]_{ij,ab} X_{ab}(t) + \sum_{ab} Q_{ij,ab} X_{ab}(t) + \sum_{abc} Q_{ja,bc} X_{ibc}(t) + \sum_{abc} Q_{ai,bc} X_{bcj}(t), \tag{3}$$

where $X_{ijk}(t)$ is the frequency (or probability) of trinucleotide ijk at time t, $\otimes$ is the Kronecker tensor product, and $\mathbb{I}$ is the 4×4 identity matrix. The first term accounts for single-nucleotide substitutions, the three last terms for dinucleotide substitutions. Dinucleotide substitutions can give rise to the dinucleotide ij in different ways. Of course a dinucleotide ab can substitute into the dinucleotide ij (second term). But we can also get the dinucleotide ij if a trinucleotide ibc containing a i on the first base, undergoes a substitution $bc \to ja$ and becomes a trinucleotide ija (third term). We also get the dinucleotide ij if a trinucleotide bcj containing a j on the third base, undergoes a substitution $bc \to ai$ and becomes a trinucleotide aij (fourth term).

The time evolution for the composition in dinucleotides (eq. (3)) therefore depends on the composition in trinucleotides, whose time evolution will in turn depend on the composition in quadrinucleotides, and so on. We are thus faced with an infinite hierarchy of equations [26]. This is the main mathematical difficulty of the neighbor-dependent model, as the infinite hierarchy of equations cannot be solved exactly in general.

Remark. Note that the infinite hierarchy of equations is in fact highly redundant. The composition in nucleotides can be obtained from the composition in dinucleotides, which can be obtained from the composition in trinucleotides, and so on

$$X_i = \sum_a X_{ia} = \sum_a X_{ai},$$
$$X_{ij} = \sum_a X_{ija} = \sum_a X_{aij}, \quad \ldots . \tag{4}$$

Using eq. (4) between compositions in nucleotides and dinucleotides and the time evolution eq. (3) for the composition in dinucleotides, one recovers the time evolution

eq. (2) for the composition in nucleotides. Similarly, the time evolution for composition in n-nucleotides implies all the time evolutions for lower numbers of nucleotides through relations like eq. (4).

2.1.2 The two-cluster approximation

To our knowledge, only Bérard *et al.* [29] succeeded in proving exact results regarding the neighbor-dependent model. The authors in [26] proposed instead to truncate the infinite hierarchy using the *two-cluster approximation*:

$$X_{ijk} \simeq \frac{X_{ij}X_{jk}}{X_j}\,. \tag{5}$$

This approximation is equivalent to state that the sequence is a first-order Markov chain (in genomic position). Then eq. (3) is closed, with trinucleotide frequencies given by eq. (5). In further numerical examples, we will explicitly use the two-cluster approximation to compute the time evolution of the dinucleotide frequencies. However, the perturbative resolution of DNA composition evolution, that yields the decomposition of the skew into transcription and replication associated components, does not rely on this approximation.

2.1.3 Odds ratios (observed over expected values)

When there are no neighbor-dependent substitution rates, *i.e.* when $Q = 0$, the solution of eq. (3) is $X_{ij}(t) = X_i(t)X_j(t)$. The observed frequencies of dinucleotides X_{ij} are then equal to their expected value X_iX_j. A way to quantify the presence of neighbor-dependence consists in computing the so-called odds ratios (or observed over expected values)

$$\rho_{ij} = \frac{X_{ij}}{X_iX_i}\,. \tag{6}$$

Odds ratios clearly different from 1 have been an indication, directly derived from the sequence, that neighbor dependence holds in most genomes [30].

2.2 Extending parity rules type 1 (PR1) and type 2 (PR2) to the neighbor-dependent model

The PR1 [31] and PR2 [31,32] strand symmetries defined in paper I for neighbor-independent and mononucleotide frequencies respectively, can be extended to neighbor-dependent substitution rates and dinucleotide frequencies. Let us recall that the reverse complement of a dinucleotide ij is the dinucleotide j^ci^c. Hence four dinucleotides are their own reverse complement

$$\begin{aligned}(TA)^c &= TA, \qquad (AT)^c = AT,\\ (GC)^c &= GC, \qquad (CG)^c = CG.\end{aligned} \tag{7}$$

The dinucleotide frequencies computed on the complementary strand are given by reverse complementarity

$$X^c_{ij} = X_{(ij)^c} = X_{j^ci^c}. \tag{8}$$

We note that the CG frequency is strand symmetric: $X^c_{CG} = X_{CG}$. The neighbor-dependent substitution rate matrix computed on the complementary strand is also given by reverse complementarity

$$Q^c_{ij,ab} = Q_{(ij)^c,(ab)^c} = Q_{j^ci^c,b^ca^c}. \tag{9}$$

We can decompose Q into symmetrical and asymmetrical parts under strand exchange symmetry, $Q = Q^s + Q^a$ with

$$Q^s = \frac{Q + Q^c}{2}\,, \qquad Q^a = \frac{Q - Q^c}{2}\,. \tag{10}$$

In the neighbor-dependent model, PR2 extends to the dinucleotides frequencies. Under symmetrical substitution rates (PR1), the frequencies of reverse complementary dinucleotides are equal at equilibrium (PR2)

$$\text{if} \quad M = M^s, \quad Q = Q^s \quad \text{then} \quad X^*_{ij} = X^*_{j^ci^c}. \tag{11}$$

3 Focusing on the CpG → TpG substitutions

3.1 Substitutional asymmetry

As detailed in appendix B of paper I, substitutions were tabulated in the human lineage since its divergence with chimpanzee using macaca and orangutan as outgroups [33]. Sequences were divided into CpG and non-CpG sites in the ancestral human-chimpanzee genome (CpG means a C followed by a G in the DNA sequence *i.e.* 5′-CG-3′). Cytosine when methylated can spontaneously deaminate into thymine. In vertebrates genomes, most CpG dinucleotides have their cytosine methylated with the exception of specific genomic regions called CpG islands [34]. As a result the CpG dinucleotide is hypermutable, and the $CpG \to TpG$ and its reverse complement $CpG \to CpA$ are by far the principal neighbor-dependent substitutions rates [24,27]. The twelve neighbor-independent substitution rates were determined for non-CpG sites. The two neighbor-dependent $r = CpG \to TpG$ and $r^c = CpG \to CpA$ substitution rates were determined for CpG sites [21]. CpG islands and exons were excluded from the analysis as they are unlikely to evolve neutrally. The first and last 500 bp of intronic sequences were also excluded to avoid bias due to splicing sites [12]. Substitution rates were computed separately in genic (+), genic (−), and intergenic regions of given replication fork polarity estimated from the derivative of the mean replication timing (MRT) as explained in the appendix A of paper I. As a substitute to germline replication fork polarity, we used the replication fork polarity determined from the MRT obtained in the HeLa cell line [33]: $p(x) \simeq vT_s\mathrm{dMRT}/\mathrm{d}x$ where the replication fork velocity $v = 0.64\,\mathrm{kbp/min}$ has been measured by DNA combing and where the S-phase duration was estimated to be $T_S \simeq 7\,\mathrm{h}$ [35].

As shown in fig. 1, where the $(CpG \to TpG)^a$ asymmetry is plotted *versus* the replication fork polarity, PR1 is not only broken in genic regions but also in intergenic regions. Actually this substitutional asymmetry decomposes

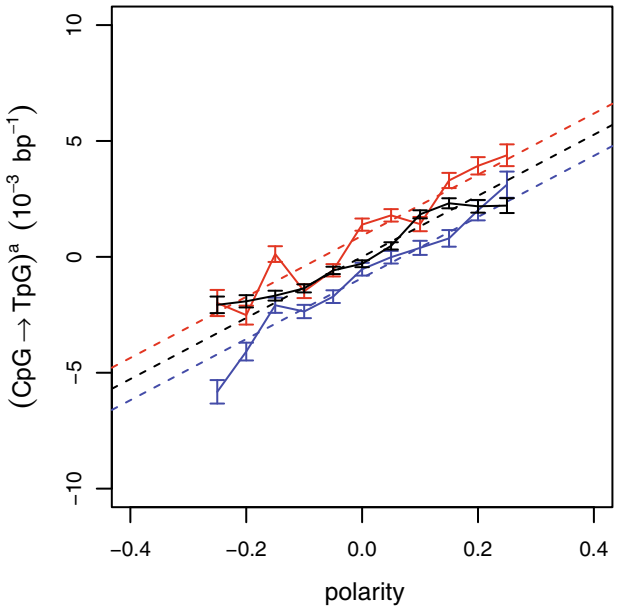

Fig. 1. $(CpG \to TpG)^a$ substitutional asymmetry *versus* replication fork polarity (determined in HeLa cell line) in genic sense (red), intergenic (black) and genic antisense (blue) regions. Substitution rates, replication fork polarity and gene orientation were computed on the reference strand.

into transcription- and replication-associated components consistently with the minimal model for substitutional asymmetry proposed in paper I (sect. 3):

$$r^a = \begin{cases} pr_R^a + r_T^a, & \text{genic } (+), \\ pr_R^a, & \text{intergenic}, \\ pr_R^a - r_T^a, & \text{genic } (-). \end{cases} \tag{12}$$

The coefficients estimated by least-squares fits of the data in fig. 1 confirm the existence of a strong replication-associated asymmetry

$$r_R^a = 13.179 \pm 0.734, \tag{13}$$

as compared to the very weak transcription-associated asymmetry

$$r_T^a = 0.912 \pm 0.142. \tag{14}$$

Furthermore the $(CpG \to TpG)^a$ asymmetry correlates strongly with the replication fork polarity in intergenic regions ($R = 0.39$[1]), even though the replication fork polarity was determined in HeLa cells and not in the germline. For the sake of completeness, we obtained the following estimates of the symmetrical substitution rates (that verify PR1 by definition):

$$r_0^s + r_R^s = 49.207 \pm 0.337 \tag{15}$$

and

$$r_T^s = -1.901 \pm 0.413. \tag{16}$$

[1] The Pearson correlation (R value) of the $(CpG \to TpG)^a$ asymmetry with the replication fork polarity p was calculated in non-overlapping 1 Mbp windows genome wide. Only 1 Mbp windows containing at least 1 kbp of aligned CpG sites were retained ($N = 412$). The p-value is $< 10^{-15}$.

3.2 Neighbor-dependent model

3.2.1 Numerical test of neighbor dependence

Taking into account $CpG \to TpG$ substitutions only amounts to taking all the elements of the matrix Q null except [26]

$$\begin{aligned} Q_{TG,CG} &= r = r^s + r^a, \qquad Q_{CA,CG} = r^c = r^s - r^a, \\ Q_{CG,CG} &= -(r + r^c) = -2r^s. \end{aligned} \tag{17}$$

The evolution of the composition then simplifies to

$$\frac{\mathrm{d}}{\mathrm{d}t}X(t) = MX(t) + X_{CG}(t) \begin{pmatrix} r^s + r^a \\ r^s - r^a \\ -r^s + r^a \\ -r^s - r^a \end{pmatrix} \tag{18}$$

in the $\{T, A, G, C\}$ coordinates. Therefore we recover the time evolution equation of the neighbor-independent model (sect. 4.1 of paper I) with an additional source term that depends on the CG frequency.

As reported in fig. 2, to illustrate the properties of neighbor dependence in the human genome, we investigated observed dinucleotide frequencies X_{ij} *versus* expected dinucleotide frequencies $X_i X_j$ at equilibrium for four different models. The dinucleotide frequencies at equilibrium were determined by integrating numerically the differential eq. (3) with the two-cluster approximation (eq. (5)). The four models correspond to special cases of the model eq. (18), with or without neighbor dependence, and evolving under PR1 or PR1 breaking. For the two neighbor-independent models ($r = 0$ in figs. 2A and B), observed dinucleotide frequencies are equal to their expected values. On the opposite, for the neighbor-dependent models ($r \neq 0$ in figs. 2C and D) the odds ratios are clearly different from 1. As expected the odds ratio of the CG dinucleotide is decreased, whereas the odds ratios of the TG and CA dinucleotides are increased. For models under PR1 ($M = M^s, r = r^s$ in figs. 2A and C), observed frequencies of reverse complementary dinucleotides are equal, as their expected values. The composition does satisfy PR2 ($[G] = [C]$ and $[T] = [A]$), as verified in figs. 2A and C, the values are clustered into three groups along the x-axis, corresponding to the only three different expected dinucleotide frequencies $[X][Y]$ values ($[G][G]$, $[T][T]$ or $[G][T]$). Furthermore, for the neighbor-dependent model under PR1 (fig. 2C), the observed dinucleotide frequencies are not equal to their expected values, but the observed frequencies of reverse complementary dinucleotides are nonetheless equal. For instance in fig. 2C, TG and CA observed frequencies are equal $[TG] = [(TG)^c] = [CA]$ whereas the CG observed frequency is only equal to itself $[CG] = [(CG)^c]$. Of course when the PR1 symmetry is broken (figs. 2B and D), the PR2 symmetry is broken for both expected and observed dinucleotide frequencies.

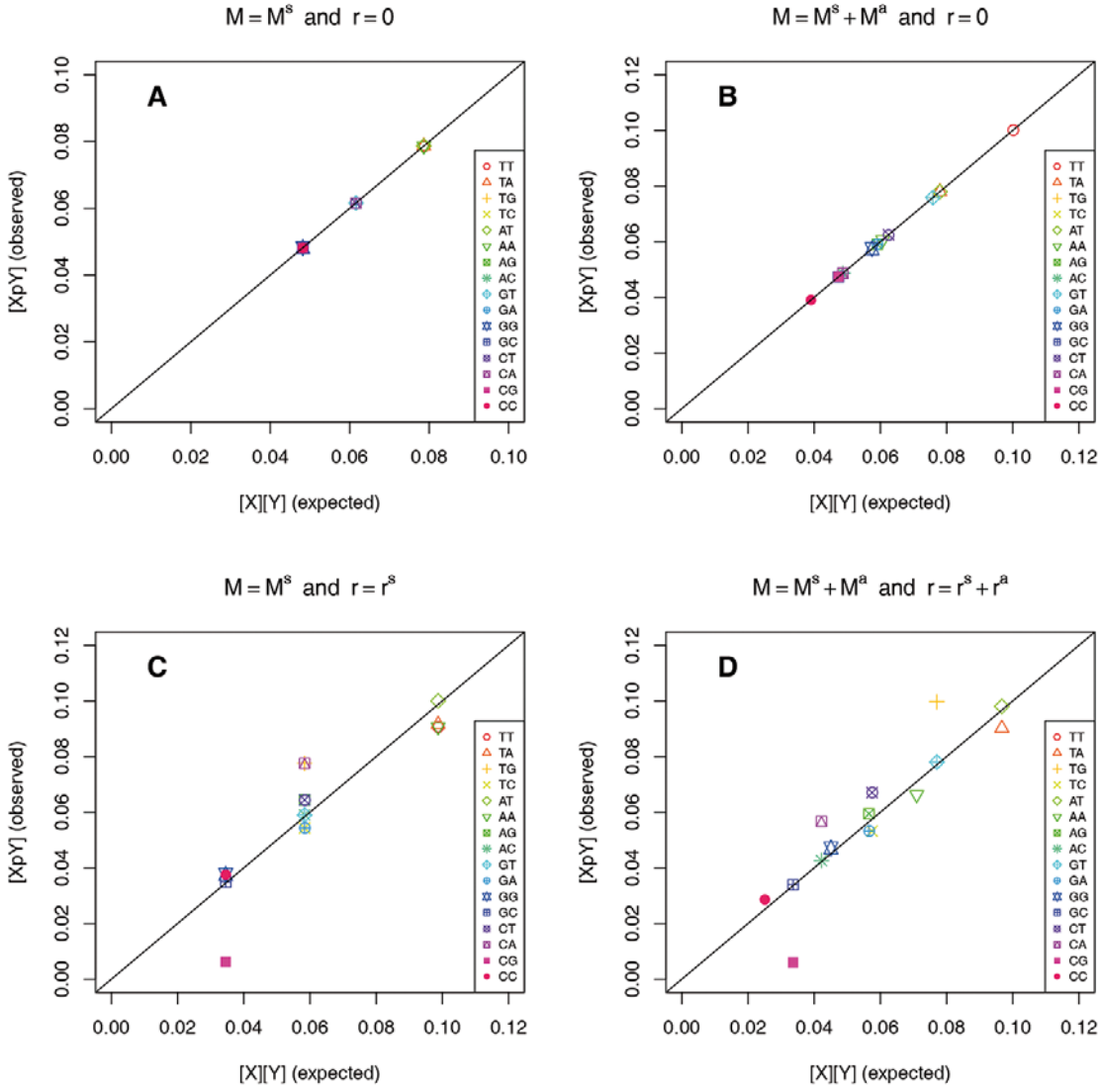

Fig. 2. Observed and expected dinucleotide frequencies at equilibrium. (A) Neighbor-independent model ($r = 0$) with symmetrical substitution rates ($M = M^s$). (B) Neighbor-independent model ($r = 0$) with substitutional asymmetry ($M^a \neq 0$). (C) Neighbor-dependent model ($r \neq 0$) with symmetrical substitution rates ($M = M^s$, $r = r^s$). (D) Neighbor-dependent model ($r \neq 0$) with substitutional asymmetry ($M^a \neq 0$, $r^a \neq 0$).

3.2.2 Time evolution of the skew

As emphasized in paper I (sect. 4.1), it is more convenient to consider the evolution of the DNA composition in the $\{\theta_{TA}, \theta_{GC}, S_{TA}, S_{GC}\}$ coordinates [36]

$$Y = \begin{pmatrix} \theta \\ S \end{pmatrix} = \begin{pmatrix} \theta_{TA} \\ \theta_{GC} \\ S_{TA} \\ S_{GC} \end{pmatrix} = UX = \begin{pmatrix} X_T + X_A \\ X_G + X_C \\ X_T - X_A \\ X_G - X_C \end{pmatrix}, \qquad (19)$$

where S_{TA} and S_{GC} are the compositional skews and θ_{TA} and θ_{GC} are the $T + A$ and $G + C$ contents. The $T + A$ and $G + C$ contents are invariant under strand exchange symmetry, whereas the compositional skews change sign. The change of coordinate matrix U and its inverse are given by

$$U = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \qquad U^{-1} = \frac{1}{2} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & -1 \end{pmatrix}. \qquad (20)$$

It is easy to get the evolution of Y through a linear transformation of eq. (18)

$$\frac{\mathrm{d}}{\mathrm{d}t} Y(t) = NY(t) + X_{CG}(t) \begin{pmatrix} 2r^s \\ -2r^s \\ 2r^a \\ 2r^a \end{pmatrix}, \qquad (21)$$

where $N = UMU^{-1} = N^s + N^a$ with

$$N^s = UM^sU^{-1} = \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}, \tag{22}$$

$$N^a = UM^aU^{-1} = \begin{pmatrix} 0 & B \\ C & 0 \end{pmatrix}. \tag{23}$$

The A and D matrices are invariant under strand exchange symmetry whereas the B and C matrices change sign. We refer the reader to sect. 4.1.1 of paper I where the 2×2 matrices A, B, C and D are defined together with the spectral properties of A and D that are needed for the time evolution of the composition. We recall that the matrix A has one eigenvalue equal to zero associated with the eigenvector θ_A that characterizes the equilibrium composition under PR1

$$A\theta_A = 0 \tag{24}$$

and another one real strictly negative

$$A\begin{pmatrix} 1 \\ -1 \end{pmatrix} = -\frac{1}{\lambda_A}\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad \text{with} \qquad \lambda_A > 0. \tag{25}$$

The two eigenvalues of D ($k = 1, 2$) have a strictly negative real part

$$DS^{(k)} = \left[-\frac{1}{\lambda_D^{(k)}} + i\omega^{(k)}\right] S^{(k)}, \qquad \text{with} \qquad \lambda_D^{(k)} > 0. \tag{26}$$

Hence, D is invertible and $e^{tD} \to 0$ when $t \to \infty$. Actually λ_A and $\lambda_D^{(1,2)}$ are characteristic time scales of the DNA composition evolution, when the substitution rate matrix satisfies PR1.

In eq. (21), we recover the time evolution of the neighbor-independent model with an additional CG frequency dependent source term. The composition at equilibrium $\mathrm{d}Y^*/\mathrm{d}t = 0$ is then given by

$$NY^* + X^*_{CG}\begin{pmatrix} 2r^s \\ -2r^s \\ 2r^a \\ 2r^a \end{pmatrix} = 0 \quad \text{and} \quad \theta^*_{TA} + \theta^*_{GC} = 1. \tag{27}$$

Therefore a source term depending on the equilibrium CG frequency is also added to the equilibrium composition (eq. (27) of paper I).

3.3 Perturbative resolution of the neighbor-dependent model

3.3.1 Perturbative analysis

Along the line of the perturbative analyses performed in sect. 4.3.2 of paper I for the neighbor-independent model, we will consider the symmetrical parts M^s and r^s of $O(1)$, while the asymmetric parts M^a and r^a are considered $O(\epsilon)$. If we start with initial null skews $S(t_0) = 0$, the perturbative resolution of eq. (21) gives the following time evolutions:

$$\theta(t) = e^{A(t-t_0)}\,\theta(t_0) + \int_{t_0}^{t} \mathrm{d}u\, e^{A(t-u)}\, X_{CG}(u)\begin{pmatrix} 2r^s \\ -2r^s \end{pmatrix} + O(\epsilon^2), \tag{28}$$

$$S(t) = \int_{t_0}^{t} \mathrm{d}u\, e^{D(t-u)}\, C\, e^{A(u-t_0)}\,\theta(t_0) + \int_{t_0}^{t} \mathrm{d}u\, e^{D(t-u)}\, C \int_{t_0}^{u} \mathrm{d}v\, e^{A(u-v)}\, X_{CG}(v)\begin{pmatrix} 2r^s \\ -2r^s \end{pmatrix} + \int_{t_0}^{t} \mathrm{d}u\, e^{D(t-u)}\, X_{CG}(u)\begin{pmatrix} 2r^a \\ 2r^a \end{pmatrix} + O(\epsilon^2). \tag{29}$$

The perturbative resolution of eq. (27) yields the equilibrium values

$$\theta^* = \theta_A + \lambda_A X^*_{CG}\begin{pmatrix} 2r^s \\ -2r^s \end{pmatrix} + O(\epsilon^2), \tag{30}$$

$$S^* = -D^{-1}\left\{C\theta_A + C\lambda_A X^*_{CG}\begin{pmatrix} 2r^s \\ -2r^s \end{pmatrix} + X^*_{CG}\begin{pmatrix} 2r^a \\ 2r^a \end{pmatrix}\right\} + O(\epsilon^2). \tag{31}$$

We note that both the $G + C$ content and the skews are affected by the neighbor-dependent rate r.

3.3.2 Impact on the $T + A$ and $G + C$ contents

We assume that the substitution rates, and in particular the $r = (CpG \to TpG)$ substitution rate and the substitution rate matrix M, follow the minimal model equations

$$\tau^a[p, \alpha, (\pm)] = p\tau^a_R \pm \tau^a_T[\alpha], \tag{32}$$

$$\tau^s[p, \alpha, (\pm)] = \tau^s_0 + \tau^s_R + \tau^s_T[\alpha], \tag{33}$$

where p, α and $(\pm)$ correspond to the replication fork polarity, transcription rate and gene orientation respectively (see sect. 2 of paper I). We recall that the coefficients τ^a_R, τ^a_T and τ^s_T, as estimated in the human genome, were found to be small compared to the symmetrical substitution rate $\tau^s_0 + \tau^s_R$ (see the single-nucleotide substitution rate matrices reported in appendix B of paper I, or eqs. (13) to (16) for the neighbor-dependent substitution rate r). This justifies the use of perturbation theory to solve the DNA composition evolution.

The perturbative resolution of eq. (21) gives the following time evolutions of the $T + A$ and $G + C$ content:

$$\theta[p, \alpha, (\pm)](t) = \tilde{\theta}_0(t) + \theta_T[\alpha](t) + O(\epsilon^2), \tag{34}$$

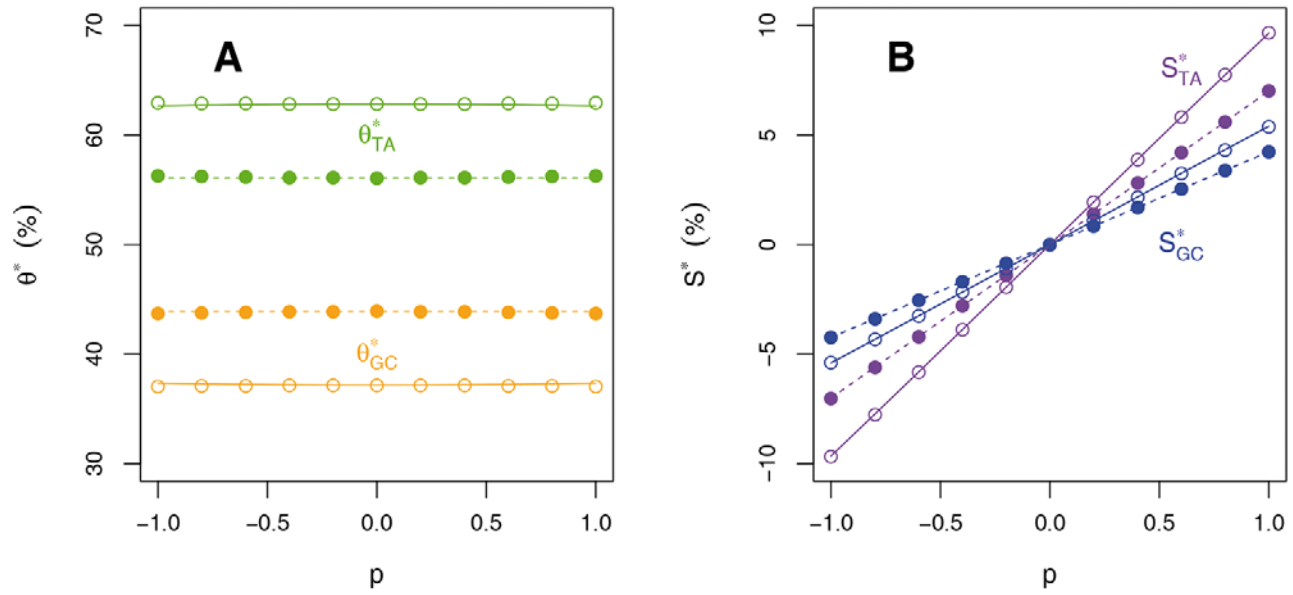

Fig. 3. Equilibrium $T+A$ and $G+C$ contents (A) and TA and GC skews (B) *versus* p. Comparison of the exact (circles, dots) and perturbative (solid line, dashed line) solutions for the neighbor-dependent (circles, solid line) and neighbor-independent (dots, dashed line) models where the substitution rate matrix $M[p] = M_0^s + M_R^s + pM_R^a$ and the neighbor-dependent substitution rate $r[p] = r_0^s + pr_R^a$ follow the minimal model (eqs. (32) and (33)) in intergenic regions of polarity p.

where

$$\tilde{\theta}_0(t) = e^{[A_0+A_R](t-t_0)}\,\theta(t_0) + \int_{t_0}^{t} \mathrm{d}u\, e^{[A_0+A_R](t-u)} X_{CG}(u) \begin{pmatrix} 2(r_0^s+r_R^s) \\ -2(r_0^s+r_R^s) \end{pmatrix}, \tag{35}$$

$$\theta_T[\alpha](t) = \int_{t_0}^{t} \mathrm{d}u\, e^{[A_0+A_R](t-u)}\, A_T[\alpha]\, \tilde{\theta}_0(u) + \int_{t_0}^{t} \mathrm{d}u\, e^{[A_0+A_R](t-u)}\, X_{CG}(u) \begin{pmatrix} 2r_T^s[\alpha] \\ -2r_T^s[\alpha] \end{pmatrix}. \tag{36}$$

The perturbative resolution of eq. (27) gives the following equilibrium $T+A$ and $G+C$ contents:

$$\theta^*[p,\alpha,(\pm)] = \tilde{\theta}_0^* + \theta_T^*[\alpha] + O(\epsilon^2), \tag{37}$$

where

$$\tilde{\theta}_0^* = \theta_{[A_0+A_R]} + \lambda_{[A_0+A_R]}\, X_{CG}^* \begin{pmatrix} 2(r_0^s+r_R^s) \\ -2(r_0^s+r_R^s) \end{pmatrix}, \tag{38}$$

$$\theta_T^*[\alpha] = \lambda_{[A_0+A_R]} \left\{ A_T[\alpha]\, \tilde{\theta}_0^* + X_{CG}^* \begin{pmatrix} 2r_T^s[\alpha] \\ -2r_T^s[\alpha] \end{pmatrix} \right\}. \tag{39}$$

As compared to the neighbor-independent model (see sect. 4.3.2 in paper I), we note that the neighbor-dependent rate r impacts on both the $\tilde{\theta}_0$ and $\theta_T[\alpha]$ coefficients in eqs. (34) to (39). As expected the $G+C$ content still does not depend on the replication fork polarity or gene orientation (fig. 3A). Indeed it depends on all the variables that determine the symmetrical substitution rates including the neighbor-dependent ones.

3.3.3 The skews still decompose into transcription- and replication-associated components

The perturbative resolution of eq. (21), with initial null skews $S(t_0) = 0$, yields the following time evolution of the TA and GC skews:

$$S[p,\alpha,(\pm)](t) = pS_R(t) \pm S_T[\alpha](t) + O(\epsilon^2), \tag{40}$$

where

$$S_R(t) = \int_{t_0}^{t} \mathrm{d}u\, e^{[D_0+D_R](t-u)}\, C_R\, \tilde{\theta}_0(u) + \int_{t_0}^{t} \mathrm{d}u\, e^{[D_0+D_R](t-u)}\, X_{CG}(v) \begin{pmatrix} 2r_R^a \\ 2r_R^a \end{pmatrix}, \tag{41}$$

$$S_T[\alpha](t) = \int_{t_0}^{t} \mathrm{d}u\, e^{[D_0+D_R](t-u)}\, C_T[\alpha]\, \tilde{\theta}_0(u) + \int_{t_0}^{t} \mathrm{d}u\, e^{[D_0+D_R](t-u)}\, X_{CG}(v) \begin{pmatrix} 2r_T^a[\alpha] \\ 2r_T^a[\alpha] \end{pmatrix}. \tag{42}$$

The perturbative resolution of eq. (27) gives the equilibrium TA and GC skew values

$$S^*[p,\alpha,(\pm)] = pS_R^* \pm S_T^*[\alpha] + O(\epsilon^2), \tag{43}$$

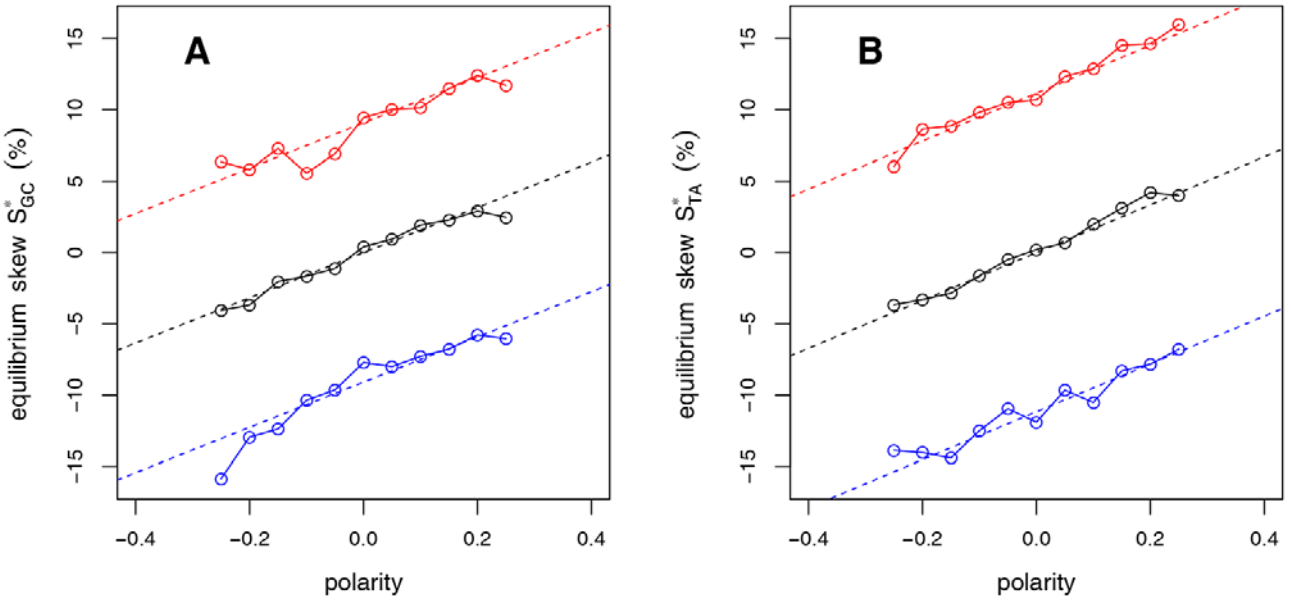

Fig. 4. The compositional asymmetry decomposes into transcription- and replication-associated components. Equilibrium skews S^*_{GC} (A) and S^*_{TA} (B) *versus* the replication fork polarity (see text and paper I) in genic sense (red), intergenic (black) and genic antisense (blue) regions. The equilibrium skews were computed taking into account the neighbor-dependent $CpG \to TpG$ substitution rate (sect. 3.2). Substitution rates, equilibrium skews, replication fork polarity and gene orientation were computed on the reference strand. Dashed lines correspond to the least-squares fit to a line following the linear model eq. (43). The so-obtained linear regression coefficients are reported in table 1.

Table 1. Transcription- and replication-associated compositional asymmetries. Coefficients $S^*_{GC,T}, S^*_{TA,T}, S^*_{GC,R}$ and $S^*_{TA,R}$ (%) of the model eq. (43), obtained by least-squares fits to a line in fig. 4 ($r \neq 0$) together with the corresponding values obtained with the neighbor-independent model in paper I ($r = 0$).

Model, S^*	$S^*_{GC,T}$	$S^*_{TA,T}$	$S^*_{GC,R}$	$S^*_{TA,R}$
Neighbor-dependent ($r \neq 0$)	9.08 ± 0.19	11.16 ± 0.13	15.84 ± 0.99	16.77 ± 0.67
Neighbor-independent ($r = 0$)	7.02 ± 0.16	10.8 ± 0.16	10.54 ± 0.82	13.64 ± 0.85

where

$$S^*_R = -[D_0 + D_R]^{-1} \left\{ C_R \tilde{\theta}^*_0 + X^*_{CG} \begin{pmatrix} 2r^a_R \\ 2r^a_R \end{pmatrix} \right\}, \qquad (44)$$

$$S^*_T[\alpha] = -[D_0 + D_R]^{-1} \left\{ C_T[\alpha] \tilde{\theta}^*_0 + X^*_{CG} \begin{pmatrix} 2r^a_T[\alpha] \\ 2r^a_T[\alpha] \end{pmatrix} \right\}. \qquad (45)$$

Therefore we recover for the compositional asymmetry the same decomposition as for the substitutional asymmetry eq. (12). From the comparison with the corresponding equations of the neighbor-independent model (sect. 4.3.2 of paper I), we note that the neighbor-dependent rate r also impacts on the S_R and $S_T[\alpha]$ coefficients.

As shown in fig. 3, when using the substitution rates determined in intergenic regions of replication fork polarity p, the perturbative solution (eqs. (37) and (43)) for the equilibrium values are indistinguishable from the exact solutions of eq. (27). As predicted by eq. (43) the equilibrium skews S^*_{TA} and S^*_{GC} are proportional to p (fig. 3B). In contrast, as predicted by eq. (37), the $G+C$ and $T+A$ contents at equilibrium do not depend upon p (fig. 3A). As shown in fig. 3, when comparing these values to the ones previously obtained in the neighbor-independent model ($r = 0$) in paper I, we confirm that the neighbor-dependent rate r does impact significantly on both the skews and the $G+C$ content. Furthermore, as reported in fig. 4, when computed in genic sense, intergenic and genic antisense regions, the equilibrium skews S^*_{GC} and S^*_{TA} both decompose into transcription- and replication-associated components consistently with the formal derivation eq. (43). The coefficients S^*_T and S^*_R estimated by least-squares fits to a line (dashed lines in fig. 4) are reported in table 1 together with the corresponding values obtained with the neighbor-independent model in paper I. We see that the additive contribution of the neighbor-dependent term significantly strengthens the coefficients S^*_T and S^*_R without changing their sign. $S^*_{TA,T}$ and $S^*_{GC,T}$ asymptotic skews associated with transcription are found positive, as well as $S^*_{TA,R}$ and $S^*_{GC,R}$ skews associated with replication, in agreement with previous analyses [11,12,16,17]. As reported in table 2, these results were confirmed when computing the Pearson correlation of the equilibrium skews with the replication fork polarity p even though the replication fork polarity was

Table 2. The equilibrium asymmetry correlates with the replication fork polarity. Pearson correlation (R values) of S^*_{GC} and S^*_{TA} with the replication fork polarity p calculated in non-overlapping 1 Mbp windows genome wide. Only 1 Mbp windows containing at least 100 kbp of aligned (intergenic) sequences and at least 100 kbp of repeat-masked (intergenic) sequences were retained for the neighbor-independent case ($N = 1982$). Only 1 Mbp windows that further contain at least 1 kbp of aligned CpG sites were retained for the neighbor-dependent case ($N = 412$). All p-values are $< 10^{-12}$.

Model, S^*	S^*_{GC}	S^*_{TA}
Neighbor-dependent ($r \neq 0$)	0.34	0.42
Neighbor-independent ($r = 0$)	0.22	0.30

determined in the HeLa cells and not in the germline (see paper I).

4 Time dependence of substitution rates

Substitutional patterns have been also shown to depend on time [23,25]. In this case all substitution rates (and all the parameters derived from them) depend on time. Importantly, on the contrary to the time homogeneous case, the composition does not necessarily converge towards an equilibrium value. Nonetheless, most results established in the time homogeneous case generalize to the time-dependent case. For instance, PR2 is still verified if the substitutional pattern satisfies PR1 [36], and the skew can still be decomposed into transcription- and replication-associated components if the substitution rates follow the minimal model (eqs. (32) and (33)).

4.1 Neighbor-independent model

4.1.1 Equilibrium composition interpreted as the current direction of evolution

If we take into account the time dependence of the substitution rates, the neighbor-independent model evolution is governed by the equation

$$\frac{\mathrm{d}}{\mathrm{d}t}X(t) = M(t)X(t). \tag{46}$$

In general the solution of eq. (46) does not converge. At each time t, the composition starts converging towards the equilibrium value $X^*(t)$, defined from the matrix $M(t)$. As the equilibrium composition $X^*(t)$ changes over time, the composition $X(t)$ may never actually reach an equilibrium state. In this perspective the equilibrium composition $X^*(t)$ gives the current direction of evolution, not the long-term asymptotic value (that may even not exist) of the composition. With this interpretation in mind, the perturbative solutions for the equilibrium $G + C$ content (eq. (71) of paper I) and the skews (eq. (65) of paper I) are still valid, but they of course depend explicitly on time.

4.1.2 PR2 is still valid for time-dependent and strand-symmetric substitution rates

PR2 was first proved [37] for the equilibrium composition for time-independent substitution rates (sect. 2.1 of paper I). But if the composition never reaches the equilibrium state, are we certain that PR2 is still satisfied? Under symmetrical substitution rates (PR1), the evolution of the skews decouples from the evolution of the $T+A$ and $G+C$ contents [36]

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = A(t)\theta(t), \tag{47}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}S(t) = D(t)S(t). \tag{48}$$

Lobry and Lobry [36] showed that the $G + C$ content on one side and the skews on the other have distinct long term behavior. The $G+C$ content generally never reaches equilibrium, $(G + C)^*(t)$ only gives the ever changing direction of evolution. On the opposite the skews always decay towards zero, as $S^*(t) = 0$ gives at all times the same direction of evolution. Therefore whatever the time-dependent substitutional pattern, under symmetrical substitution rates (PR1), the nucleotide composition will always satisfy PR2 asymptotically.

4.2 Neighbor-dependent model

4.2.1 Solving time-dependent differential equations

If we take into account the time dependence of the substitution rates, the neighbor-dependent model eq. (21) may be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}Y(t) = N(t)Y(t) + X_{CG}(t)\begin{pmatrix} 2r^s(t) \\ -2r^s(t) \\ 2r^a(t) \\ 2r^a(t) \end{pmatrix} \tag{49}$$

in the $\{\theta_{TA}, \theta_{GC}, S_{TA}, S_{GC}\}$ coordinates. This model falls into the class of time-dependent linear differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t}Y(t) = N(t)Y(t) + Z(t), \tag{50}$$

where $Z(t)$ is a source term. When solving perturbatively eq. (49), we systematically encounter differential equations of this class. To solve eq. (50) we need to introduce the time-ordered exponential [38]

$$Te^{\int_{t_0}^{t} \mathrm{d}uN(u)} = \mathbb{I} + \sum_{n\geq 1} \int_{t\geq t_n\geq \ldots \geq t_1 \geq t_0} \mathrm{d}t_n \ldots \mathrm{d}t_1 N(t_n)\ldots N(t_1), \tag{51}$$

where $\mathbb{I}$ is the identity matrix. Importantly the time-ordered exponential satisfies the following properties:

$$\frac{\mathrm{d}}{\mathrm{d}t}\, Te^{\int_{t_0}^{t} \mathrm{d}uN(u)} = N(t)\, Te^{\int_{t_0}^{t} \mathrm{d}uN(u)} \tag{52}$$

and

$$Te^{\int_{t_0}^{t} \mathrm{d}uN(u)} \Big|_{t=t_0} = \mathbb{I}. \tag{53}$$

The solution of eq. (50) with the initial condition $Y(t_0)$ at t_0 is given by

$$Y(t) = Te^{\int_{t_0}^{t} \mathrm{d}uN(u)} Y(t_0) + \int_{t_0}^{t} \mathrm{d}u\, Te^{\int_{u}^{t} \mathrm{d}vN(v)} Z(u). \tag{54}$$

Indeed the solution given by eq. (54) satisfies both eq. (50) and the initial condition thanks to the properties of the time-ordered exponential (eqs. (52) and (53)). Besides uniqueness of the solution is ensured by the Cauchy-Lipschitz theorem.

We recall that the solution of differential eq. (50) in the time-independent case $N(t) = N$ is given by

$$Y(t) = e^{(t-t_0)N} Y(t_0) + \int_{t_0}^{t} \mathrm{d}u\, e^{(t-u)N} Z(u). \tag{55}$$

Consistently the time-ordered exponential reduces to the ordinary exponential in the time-independent case

$$Te^{\int_{t_0}^{t} \mathrm{d}uN(u)} = e^{(t-t_0)N}, \qquad \text{when} \qquad N(t) = N. \tag{56}$$

Therefore the resolution of the time-dependent differential equation actually amounts to replace the ordinary exponential in eq. (55) by the time-ordered exponential in eq. (54).

4.2.2 Perturbative analysis of the compositional asymmetry with time-dependent substitution rates

If we take into account the time dependence of substitution rates, our minimal model writes

$$\tau^a[p, \alpha, (\pm)](t) = p\tau_R^a(t) \pm \tau_T^a[\alpha](t), \tag{57}$$

$$\tau^s[p, \alpha, (\pm)](t) = [\tau_0^s + \tau_R^s](t) + \tau_T^s[\alpha](t), \tag{58}$$

for all substitution rates τ. All the results derived in the previous perturbative analyses (sect. 4.3 of paper I and sect. 3.3) can be extended to the time-dependent case. The equilibrium skews are still given by eqs. (43) to (45), but they now depend explicitly on time. The time evolution of the skews is still given by eqs. (40) to (42), if we systematically replace ordinary exponentials by time-ordered exponentials in the expressions of $S_R(t)$ and $S_T[\alpha](t)$

$$\begin{aligned} S_R(t) = &\int_{t_0}^{t} \mathrm{d}u\, Te^{\int_{u}^{t} \mathrm{d}v[D_0+D_R](v)}\, C_R(u)\, \tilde{\theta}_0(u) \\ &+ \int_{t_0}^{t} \mathrm{d}u\, Te^{\int_{u}^{t} \mathrm{d}v[D_0+D_R](v)} \\ &\times X_{CG}(u) \begin{pmatrix} 2r_R^a(u) \\ 2r_R^a(u) \end{pmatrix}, \end{aligned} \tag{59}$$

$$\begin{aligned} S_T[\alpha](t) = &\int_{t_0}^{t} \mathrm{d}u\, Te^{\int_{u}^{t} \mathrm{d}v[D_0+D_R](v)}\, C_T[\alpha](u)\, \tilde{\theta}_0(u) \\ &+ \int_{t_0}^{t} \mathrm{d}u\, Te^{\int_{u}^{t} \mathrm{d}v[D_0+D_R](v)} \\ &\times X_{CG}(u) \begin{pmatrix} 2r_T^a[\alpha](u) \\ 2r_T^a[\alpha](u) \end{pmatrix} \end{aligned} \tag{60}$$

with

$$\begin{aligned} \tilde{\theta}_0(t) = &Te^{\int_{t_0}^{t} \mathrm{d}v[A_0+A_R](v)}\, \theta(t_0) \\ &+ \int_{t_0}^{t} \mathrm{d}u\, Te^{\int_{u}^{t} \mathrm{d}v[A_0+A_R](v)} \\ &\times X_{CG}(u) \begin{pmatrix} 2[r_0^s + r_R^s](u) \\ -2[r_0^s + r_R^s](u) \end{pmatrix}. \end{aligned} \tag{61}$$

In the time-dependent case the skews can still be decomposed into a transcription- and a replication-associated contribution. As in the time homogeneous case, the replication-associated contribution is proportional to the replication fork polarity, while the transcription-associated one increases in magnitude with transcription rate and changes sign with gene orientation.

Remark. Note that the above conclusion relies on the assumption that the parameters p, α, and gene orientation do not change with time. If for example the replication fork polarity profile along the human chromosomes changes over evolutionary time then we are no longer certain that the replication-associated skew remains proportional to p.

5 Conclusion

In this paper II, we have shown that when taking into account for neighbor dependence in our minimal modeling of neutral nucleotide substitution rates proposed in paper I [22], the main conclusion obtained with the neighbor-independent model remains valid, namely the compositional asymmetry actually decomposes into a transcription- and a replication-associated component. The former increases in magnitude with transcription rate and changes sign with gene orientation whereas the latter is proportional to the replication fork polarity. These results provide strong theoretical support to the wavelet-based multi-scale methodology developed in [19] to disentangle transcription- and replication-associated skews

in mammalian genomes in the so-called replication N-domains. The genic crenel-shaped skew associated with transcription superimposes (additive for sense genes and subtractive for antisense genes) to a N-shape skew profile associated with replication [18]. Accordingly, our genome-wide analysis confirms that the linear decrease of the replication-associated component of the skew inside N-domains likely reflects a decrease in the proportion of replication forks propagating from the left (5′) to the right (3′) N-domain extremity [39]. These N-domains borders were shown to exhibit a very peculiar gene organization and chromatin state [18,40,41]. They were proposed to be replication origins, evolutionary conserved and active in the germline [16,17,21,42]. Indeed, the existence of a new type of replication domains presenting gradients of replication fork polarity contrasts with the previously proposed dichotomic segmentation of mammalian chromosomes in regions replicated either by multiple synchronous origins with equal proportion of forks coming from both directions (0.2–2 Mbp Constant Timing Regions) or by unidirectional replication forks (0.1–0.6 Mbp Transition Timing Regions) [43–46]. Taking advantage of replication timing profile in several human cell types [33,47], we have recently shown [39] that, as the signature of the replication fork polarity, the derivative of the replication timing profile behaves as a N in so-called U-shaped replication timing domains demonstrating that replication N-domains are not specific to the germline but robustly observed in stem and somatic cell types as covering about 50% of the human genome. Furthermore as observed with "master" replication origins bordering the skew N-domains [20,40] the early initiation zones bordering the replication timing U-domains are significantly enriched in open chromatin markers as well as in insulator-binding proteins CTCF and are prone to gene activity [39]. The subsequent analysis of recent Hi-C [48] and 4C [49] data has revealed that replication timing U-domains actually correspond to self-interacting structural chromatin units [39,49]. Altogether these results enlighten the fundamental role of these "islands" of open chromatin observed at U-domains borders: at the heart of a compartmentalization of chromosomes into chromatin units of independent replication and of coordinated gene transcription, they likely are the corner stone of a highly paralleled spatio-temporal replication program in the human genome and more generally in mammalian genomes [20,39]. Altogether these results open new perspectives in the modeling of the replication program in higher eukaryotes [50,51] possibly differing from those proposed in yeast [52–54] where the existence of a replication-associated skew has been recently demonstrated [55]. In that context, a dynamic model has been recently proposed [35,39] in which replication first initiates at replication timing U-domains borders followed by a chromatin gradient-mediated succession of secondary origin activations. It will thus be essential to determine to what extent the chromatin state influences fork progression and origin activation. Finally, the present study raises the question of the stability of the spatio-temporal replication program during evolution. According to our "minimal" modeling the observed compositional skews were generated over several hundreds Myrs, *i.e.* a time period much larger the age of the mammalian radiation; furthermore these skews are far from having reached their equilibrium values. A detailed study of genome compositional asymmetries during evolution is in current progress.

This work was supported by the Agence Nationale de la Recherche (projet "REFOPOL": Spatio-temporal program of replication of the human genome, ANR 10 BLAN 1615).

References

1. J.M. Freeman, T.N. Plasterer, T.F. Smith, S.C. Mohr, Science **279**, 1827 (1998).
2. A. Beletskii, A. Grigoriev, S. Joyce, A.S. Bhagwat, J. Mol. Biol. **300**, 1057 (2000).
3. M.P. Francino, H. Ochman, Mol. Biol. Evol. **18**, 1147 (2001).
4. J.R. Lobry, Mol. Biol. Evol. **13**, 660 (1996).
5. J. Mrázek, S. Karlin, Proc. Natl. Acad. Sci. U.S.A. **95**, 3720 (1998).
6. A.C. Frank, J.R. Lobry, Gene **238**, 65 (1999).
7. E.P. Rocha, A. Danchin, A. Viari, Mol. Microbiol. **32**, 11 (1999).
8. E.R. Tillier, R.A. Collins, J. Mol. Evol. **50**, 249 (2000).
9. L. Duret, Curr. Opin. Genet. Dev. **12**, 640 (2002).
10. P. Green, B. Ewing, W. Miller, P.J. Thomas, E.D. Green, Nat. Genet. **33**, 514 (2003).
11. M. Touchon, S. Nicolay, A. Arneodo, Y. d'Aubenton-Carafa, C. Thermes, FEBS Lett. **555**, 579 (2003).
12. M. Touchon, A. Arneodo, Y. d'Aubenton-Carafa, C. Thermes, Nucleic Acids Res. **32**, 4969 (2004).
13. S. Nicolay, E.B. Brodie of Brodie, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Phys. Rev. E **75**, 032902 (2007).
14. A. Gierlik, M. Kowalczuk, P. Mackiewicz, M.R. Dudek, S. Cebrat, J. Theor. Biol. **202**, 305 (2000).
15. S. Nicolay, F. Argoul, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Phys. Rev. Lett. **93**, 108101 (2004).
16. E.B. Brodie of Brodie, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Phys. Rev. Lett. **94**, 248103 (2005).
17. M. Touchon, S. Nicolay, B. Audit, E.B. Brodie of Brodie, Y. d'Aubenton-Carafa, A. Arneodo, C. Thermes, Proc. Natl. Acad. Sci. U.S.A. **102**, 9836 (2005).
18. M. Huvet, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, A. Arneodo, C. Thermes, Genome Res. **17**, 1278 (2007).
19. A. Baker, S. Nicolay, L. Zaghloul, Y. d'Aubenton-Carafa, C. Thermes, B. Audit, A. Arneodo, Appl. Comput. Harmon. Annal. **28**, 150 (2010).

20. A. Arneodo, C. Vaillant, B. Audit, F. Argoul, Y. d'Aubenton-Carafa, C. Thermes, Phys. Rep. **498**, 45 (2011).
21. C.L. Chen, L. Duquenne, B. Audit, G. Guilbaud, A. Rappailles, A. Baker, M. Huvet, Y. d'Aubenton-Carafa, O. Hyrien, A. Arneodo *et al.*, Mol. Biol. Evol. **28**, 2327 (2011).
22. A. Baker, H. Julienne, C. Chen, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Eur. Phys. J. E **35**, 92 (2012).
23. D.G. Hwang, P. Green, Proc. Natl. Acad. Sci. U.S.A. **101**, 13994 (2004).
24. P.F. Arndt, T. Hwa, Bioinformatics **21**, 2322 (2005).
25. C.F. Mugal, H.H. von Grunberg, M. Peifer, Mol. Biol. Evol. **26**, 131 (2009).
26. P.F. Arndt, C.B. Burge, T. Hwa, J. Comput. Biol. **10**, 313 (2003).
27. S.T. Hess, J.D. Blake, R.D. Blake, J. Mol. Biol. **236**, 1022 (1994).
28. P.F. Arndt, T. Hwa, D.A. Petrov, J. Mol. Evol. **60**, 748 (2005).
29. J. Bérard, J.B. Gouéré, D. Piau, Math. Biosci. **211**, 56 (2008).
30. C. Burge, A.M. Campbell, S. Karlin, Proc. Natl. Acad. Sci. U.S.A. **89**, 1358 (1992).
31. N. Sueoka, J. Mol. Evol. **40**, 318 (1995).
32. R. Rudner, J.D. Karkas, E. Chargaff, Proc. Natl. Acad. Sci. U.S.A. **60**, 921 (1968).
33. C.L. Chen, A. Rappailles, L. Duquenne, M. Huvet, G. Guilbaud, L. Farinelli, B. Audit, Y. d'Aubenton-Carafa, A. Arneodo, O. Hyrien *et al.*, Genome Res. **20**, 447 (2010).
34. M.M. Suzuki, A. Bird, Nat. Rev. Genet. **9**, 465 (2008).
35. G. Guilbaud, A. Rappailles, A. Baker, C.L. Chen, A. Arneodo, A. Goldar, Y. d'Aubenton-Carafa, C. Thermes, B. Audit, O. Hyrien, PLoS Comput. Biol. **7**, e1002322 (2011).
36. J.R. Lobry, C. Lobry, Mol. Biol. Evol. **16**, 719 (1999).
37. J.R. Lobry, J. Mol. Evol. **40**, 326 (1995).
38. N. Van Kampen, *Stochastic Processes in Physics and Chemistry*, 3rd edition (North-Holland, Amsterdam, 2007).
39. A. Baker, B. Audit, C. Chen, B. Moindrot, A. Leleu, G. Guilbaud, A. Rappailles, C. Vaillant, A. Goldar, F. Mongelard *et al.*, PLoS Comput. Biol. **8**, e1002443 (2012).
40. B. Audit, L. Zaghloul, C. Vaillant, G. Chevereau, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Nucleic Acids Res. **37**, 6064 (2009).
41. C. Lemaitre, L. Zaghloul, M.F. Sagot, C. Gautier, A. Arneodo, E. Tannier, B. Audit, BMC Genomics **10**, 335 (2009).
42. B. Audit, S. Nicolay, M. Huvet, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Phys. Rev. Lett. **99**, 248102 (2007).
43. S. Farkash-Amar, D. Lipson, A. Polten, A. Goren, C. Helmstetter, Z. Yakhini, I. Simon, Genome Res. **18**, 1562 (2008).
44. I. Hiratani, T. Ryba, M. Itoh, T. Yokochi, M. Schwaiger, C.W. Chang, Y. Lyou, T.M. Townes, D. Schubeler, D.M. Gilbert, PLoS Biol. **6**, e245 (2008).
45. R. Desprat, D. Thierry-Mieg, N. Lailler, J. Lajugie, C. Schildkraut, J. Thierry-Mieg, E.E. Bouhassira, Genome Res. **19**, 2288 (2009).
46. T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T.C. Schulz, A.J. Robins, S. Dalton, D.M. Gilbert, Genome Res. **20**, 761 (2010).
47. R.S. Hansen, S. Thomas, R. Sandstrom, T.K. Canfield, R.E. Thurman, M. Weaver, M.O. Dorschner, S.M. Gartler, J.A. Stamatoyannopoulos, Proc. Natl. Acad. Sci. U.S.A. **107**, 139 (2010).
48. E. Lieberman-Aiden, N.L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R. Lajoie, P.J. Sabo, M.O. Dorschner *et al.*, Science **326**, 289 (2009).
49. B. Moindrot, B. Audit, P. Klous, A. Baker, C. Thermes, W. de Laat, P. Bouvet, F. Mongelard, A. Arneodo, Nucleic Acids Res. **40**, 9470 (2012).
50. O. Hyrien, K. Marheineke, A. Goldar, Bioessays **25**, 116 (2003).
51. O. Hyrien, A. Goldar, Chromosome Res. **18**, 147 (2010).
52. A.P.S. de Moura, R. Retkute, M. Hawkins, C.A. Nieduszynski, Nucleic Acids Res. **38**, 5623 (2010).
53. S.C.H. Yang, N. Rhind, J. Bechhoefer, Mol. Syst. Biol. **6**, 404 (2010).
54. A. Baker, B. Audit, S.H. Yang, J. Bechhoefer, A. Arneodo, Phys. Rev. Lett. **108**, 268101 (2012).
55. M.C. Marsolier-Kergoat, A. Goldar, Mol. Biol. Evol. **29**, 893 (2012).