Sequence analysis

Is there an acceleration of the CpG transition rate during the mammalian radiation?

M. Peifer1,⁎, J. E. Karro2,3 and H. H. von Grünberg1

1Institute of Chemistry, Karl-Franzens University Graz, Graz, Austria, 2Department of Computer Science and Systems Analysis and 3Department of Microbiology, Miami University, OH, USA

Received on April 11, 2008; revised and accepted on July 27, 2008

Advance Access publication July 28, 2008

Associate Editor: Martin Bishop

ABSTRACT

Motivation: In this article we build a model of the CpG dinucleotide substitution rate and use it to challenge the claim that, that rate underwent a sudden mammalian-specific increase approximately 90 million years ago. The evidence supporting this hypothesis comes from the application of a model of neutral substitution rates able to account for elevated CpG dinucleotide substitution rates. With the initial goal of improving that model's accuracy, we introduced a modification enabling us to account for boundary effects arising by the truncation of the Markov field, as well as improving the optimization procedure required for estimating the substitution rates.

Results: When using this modified method to reproduce the supporting analysis, the evidence of the rate shift vanished. Our analysis suggests that the CpG-specific rate has been constant over the relevant time period and that the asserted acceleration of the CpG rate is likely an artifact of the original model.

Contact: peifer@uni-graz.at

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

In much of the bioinformatics literature addressing mammalian neutral nucleotide substitution rates, researchers make the simplifying assumption that the rate at which a given base undergoes substitution is unaffected by the content of its neighbors—that it is free of its context (Hardison et al., 2003; Hasegawa et al., 1985; Jukes and Cantor, 1969; Lio and Goldman, 1998; Tyekucheva et al., 2008). However, it is well established that this assumption only approximates the biological reality. The best known example of a context-sensitive substitution is triggered by the methylation of CpG dinucleotides in vertebrates. Spontaneous, hydrolytic deamination then converts the 5-methylcytosine into thymine, while the same process would result in a uracil were methylation not in play. Since the repair mechanism is considerably less efficient in correcting the \(C \rightarrow T\) transition than the \(C \rightarrow U\) mutation, a particularly high rate for \(CpG \rightarrow TpG\) (and \(CpG \rightarrow CpA\) on the reverse strand) is observed (Coulondre et al., 1978; Ehrlich et al., 1986; Razin and Riggs, 1980; Wiebauer et al., 1993). Estimates of the rate for this context-dependent transition ranges from 10 to 60 times the rate of the other, single-nucleotide transitions (Arndt et al., 2003a; Blake et al., 1992; Fryxell and Zuckerkandl, 2000; Hess et al., 1994; Hwang and Green, 2004; Meunier et al., 2005; Siepel and Haussler, 2004).

In a recent series of studies, Arndt and colleagues have gone to great effort to model the rate of CpG dinucleotide substitutions. To this end, they incorporate the CpG effect into their method of estimating substitution rates and then apply that method to repeat data of mammalian genomes (Arndt and Hwa, 2004, 2005; Arndt et al., 2003a, b, 2005). The results provide a picture of the behavior of the CpG substitution rate over time by relating it to the average transversion frequency. From this analysis they conclude that CpG-related substitutions occur at rates as high as 40 times the rate of the average transversions in modern mammals. Further, by applying the model to repeats of differing ages, they provide evidence that the rate of CpG substitution dramatically increased along the mammalian lineage about 90 Mya—roughly the time of the mammalian radiation (Arndt and Hwa, 2004; Arndt et al., 2003a).

Expecting to advance the results of the Arndt et al. studies, we implemented an alternative model for the analysis of context-sensitive substitution rates. We followed their methodology in that we restricted the context-dependent substitution process to a triplet of bases. However, we also tried to account for the ‘truncation errors’ that result from the consideration of triplets in isolation [as opposed to considering the whole sequence at once, as done in Hwang and Green (2004)]. In addition, we chose a different objective function used in the estimation of the substitution rates. Unexpectedly, when we fitted the boundary-corrected model to the data, the evidence for the shift in the CpG substitution rate vanished and we found a considerably higher CpG rate of substitutions in older repeats than predicted in Arndt et al. (2003a), suggesting the CpG rate has in fact been constant in time, not changing as has been suggested.

We will first compare two substitution rate models: the one used in the studies of Arndt et al., which we refer to as the simple context-dependent model (Arndt and Hwa, 2004; Arndt et al., 2003a), and our boundary-corrected model. Development of our model also requires the use of a different objective function for estimating the substitution rates than that used by Arndt et al. (Arndt and Hwa, 2004; Arndt et al., 2003a). A comparison of the estimates produced by each model will then shed light on the claim of a mammal-specific shift in CpG substitution rates and help us determine whether this shift is real.
2 METHODS

2.1 Substitution models

The technique for estimating substitution rates proposed by (Arndt et al., 2003a), and our variation on that technique, is built on Markov chain theory that has been used in numerous other studies [e.g. Jukes and Cantor (1969) Hasegawa et al. (1985); see Ewens and Grant (2005) for an overview]. The Markov-chain model is defined by a rate matrix \( \mathbf{R} \), from which we can derive \( \mathbf{P}(t) \)—a matrix quantification of the state-transitional probabilities over a time period \( t \). Each state of the chain represents one possible nucleotide configuration. When the common simplifying assumption of base independence is made (disregarding \( \mathbf{CPG} \) or other context-sensitive effects), we then have a four state Markov chain and \( \mathbf{R} \) and \( \mathbf{P} \) matrices of size \( 4 \times 4 \). If we instead account for limited contextual effect by assuming that the transitional probability of a base is dependent on its left and right neighbors, we then require a state for each possible triplet of nucleotides and the matrices are of size \( 64 \times 64 \). The resulting 64-state model has been explored in several studies (Arndt et al., 2003a; b; Lunter and Hein, 2004; Siepel and Haussler, 2004).

Previous works have shown that when looking at genome-wide average substitution rates (as opposed to local rates), we can use a model conforming to strand symmetry (Lobry and Lobry, 1999; Sueoka, 1995). That is, we can assume the rates of complementary substitutions (e.g. \( A \rightarrow C \) and \( T \rightarrow G \) ) are equal. When substitution rates are independent of context, strand symmetry implies that the model can be described by six rates: two distinct transition rates (\( r_3: A/T \rightarrow G/C \) and \( r_6: C/G \rightarrow A/T \) ), and four distinct transversion rates (which we denote \( r_1, \ldots, r_4 \) but have no need to distinguish between them in the following).

Modeling full context dependency is difficult, as it is not clear what dinucleotide substitutions exist. In non-coding regions the \( \mathbf{CPG} \) substitution is the most clearly identifiable example of a context-sensitive substitution (Siepel and Haussler, 2004). Hence, we follow the common practice of adding only this one context-dependent rate to our model (Arndt and Hwa, 2004; Arndt et al., 2003b). Specifically: for single base substitutions we continue to use our \( r_1 \) parameters defined above (e.g. \( r_5 \) represents the rate of an \( AAA \rightarrow AGA \) transition); for unique \( \mathbf{CPG} \) transitions (up to strand symmetry), we introduce a seventh parameter \( r_7 \) (representing the rate for transitions such as \( CGA \rightarrow TGA/CAA \)); it is assumed that no other substitution (e.g. \( AAA \rightarrow TTT \) ) can occur in a single step. This model, which has been the subject of numerous studies (Arndt and Hwa, 2004; Arndt et al., 2003a, b, 2005), will be referred to as the simple context-dependent model. The model structure is described in the Supplementary Materials, Equation (SB).

The simple context-dependent model is an improvement over the context-free model, but it still suffers from truncation error. Because we are considering isolated triplets, the neighbor effect is only partially corrected for in the flanking bases: while the substitution \( \text{ACG} \rightarrow \text{AGT} \) is correctly reflected by \( r_7 \), the substitution \( \text{AAC} \rightarrow \text{AAT} \) is reflected in \( r_6 \) even though the flanking \( C \rightarrow T \) transition may have been influenced by a neighboring \( G \) (Hwang and Green, 2004). In the Supplementary Materials, we describe a strategy for dealing with this boundary effect. In short, we add six new parameters \( r_{b1}, \ldots, r_{b6} \) which reflect the single-base substitution rates of the flanking bases. Differences between \( r_7 \) and \( r_{b7} \) result from the truncation error, and can be used to measure the boundary effect. The derived rate matrix, given by Equation (S12) of the Supplementary Materials, will be referred to as the boundary-corrected model and is denoted \( \mathbf{R}_b \) in the rest of this study.

2.2 Estimation of substitution frequencies

Following studies such as Hardison et al. (2003), Arndt et al. (2005), Gaffney and Keightley (2005) and Karro et al. (2008), we assume interspersed repeats reflect the neutral substitution rates. For each family of repeats, the RepeatMasker tool (Smit et al., 1996–2004) provides us with the alignment of a derived ancestral sequence to each modern interspersed repeat sequence, allowing us to calculate the number of substitutions from ancestral state \((i,j,k)\) to modern state \((i',j',k')\). After discarding all low-complexity repeats, functional repeats and repeat families covering <100 kb on the modern genome, we can calculate this information for 494 repeat families, covering roughly 40% of the human genome. Note that the presence of repeats in \( \mathbf{CPG} \) islands can be safely neglected since those repeats are statistically underrepresented. Thus, they should not contribute to a significant impact on the result.

For each repeat family, denoted \( \alpha \) and inserted at time \( -t_\alpha \) (i.e. a time \( t_\alpha \) time units in the past), the number of substitutions \( N(i,j,k \rightarrow i',j',k';t_\alpha) \) is calculated from the alignments. The substitution frequencies \( \mathbf{R}_b \) and \( \mathbf{R}_b(t_\alpha) \) can then be determined by making use of the exponential relationship \( \mathbf{P}(\mathbf{R}_b(t_\alpha)) \) and \( \mathbf{P}(\mathbf{R}_b(t_\alpha)) \), respectively, as given in Equation (S9) (Supplementary Material) for each repeat family inserted at time \( -t_\alpha \). More specifically, we need to estimate substitution frequencies such that \( \mathbf{P}(\mathbf{R}_b(t_\alpha)) \) fits the observed number of substitutions \( N(i,j,k \rightarrow i',j',k';t_\alpha) \). This leads to an optimization problem for which a suitable objective function is required. Since in the following we need to reproduce the results presented in the studies of Arndt et al. (Arndt and Hwa, 2004; Arndt et al., 2003a, b, 2005), when using the context-dependent approach without boundary correction we adopt the same objective function used by Arndt et al. (Arndt et al., 2003a); Arndt and Hwa (2004)):

\[
\mathcal{L}_1(\mathbf{R}_b(t_\alpha)) = -\sum_{i,j,k} N(i,j,k \rightarrow i',j',k';t_\alpha) \times \log \left( \mathbf{P}(i,j,k \rightarrow i',j',k';\mathbf{R}_b(t_\alpha)) \right),
\]

where \( N(i,j,k \rightarrow i',j',k';t_\alpha) = \sum_{i,j,k} N(i,j,k \rightarrow i',j',k';t_\alpha) \) and in a similar fashion \( \mathbf{P}(i,j,k \rightarrow i',j',k';\mathbf{R}_b(t_\alpha)) = \sum_{i,j,k} \mathbf{P}(i,j,k \rightarrow i',j',k';\mathbf{R}_b(t_\alpha)) \) with \( \mathbf{P}(i,j,k \rightarrow i',j',k';\mathbf{R}_b(t_\alpha)) \) denoting the components of \( \mathbf{P}(\mathbf{R}_b(t_\alpha)) \). It is important to note that summing over the flanking bases \( i',k' \) within \( N \) and \( \mathbf{P} \) (represented by the dot symbol in both expressions) can be viewed as an observation function, an idea which was introduced in Arndt et al. (2003a).

When using the boundary-corrected model, the use of \( \mathcal{L}_1 \) as an objective function would make it impossible to estimate the additional set of parameters \( r_{b1}, \ldots, r_{b6} \). We therefore choose the objective function which follows directly from the principle of maximum likelihood:

\[
\mathcal{L}_2(\mathbf{R}_b(t_\alpha)) = -\sum_{i,j,k} N(i,j,k \rightarrow i',j',k';t_\alpha) \times \log \left( \mathbf{P}(i,j,k \rightarrow i',j',k';\mathbf{R}_b(t_\alpha)) \right),
\]

where again the matrix components of \( \mathbf{P}(\mathbf{R}_b(t_\alpha)) \) are denoted by \( \mathbf{P}(i,j,k \rightarrow i',j',k';\mathbf{R}_b(t_\alpha)) \). Note that the summation over the flanking bases \( i' \) and \( k' \) has been dropped. Having thus defined the objective functions, substitution frequencies \( r_{b1}, \ldots, r_{b6} \) are obtained by minimizing Equation (1) and \( r_{b7}, \ldots, r_{b12}, r_{b13}, \ldots, r_{b18} \) are estimated by minimizing Equation (2). The minimization of both \( \mathcal{L}_1 \) and \( \mathcal{L}_2 \) was performed using the same optimization algorithm employed in Arndt et al. (2003a).

2.3 Simulations

For simplicity we will refer to the method of Arndt et al. (2003a) that uses the simple context-dependent model and objective function \( \mathcal{L}_1 \) as Method 1, or M1 [see Equation (1)]; we further refer to the method using the boundary corrected model and objective function \( \mathcal{L}_2 \) as Method 2, or M2 [see Equation (2)]. To test the performance of both methods, we simulated sequences using certain predefined rates and then estimated these rates from the synthetic data using either M1 or M2. The test is successful if predefined and estimated rates agree. A sequence of \( 10^6 \) bases was generated from a random consensus sequence. Simulations were performed by mutating an ancestral sequence and then stepping forward in time by increments of \( \Delta t = 10^{-5} \). At each step, every base is given a chance to mutate with probabilities dictated by the possible seven substitution processes, with probability computed by multiplying the corresponding substitution rate with \( \Delta t \). By repeating this procedure the sequence is propagated through time.
We note that a log-scale has been used to help show that the single nucleotide boundary effect can also be seen by calculating the difference between magnitude lower for M2 than for M1. Additionally, the importance of the CpG that the estimation error of the summation over the flanking neighbors.

We conclude from this that the bias produced by M1 must be primarily due to observe that a similar bias (shown in Fig. 1) occurs when applying M1 to a trinucleotide scale, a truncation error cannot arise. This allows us to study the bias produced by M1, we simulated an ensemble of trinucleotides using the flanking bases or the truncation error is primarily responsible for the time range where M1 seems to be susceptible to biases resulting from the pre-defined value occurred (that is, the value of r7 as predicted by method 1). The symbols indicate those spots in which the difference between the method 1 estimate and the actual value is statistically significant. We note that a log-scale has been used to help show that the single nucleotide transitions are accurately estimated for both methods.

until the desired divergence (in terms of average transversion frequency) is reached. In accordance with the results presented further below, we used the following substitution rates in the simulation: \(r_1, \ldots, r_4 = 1\), \(r_5 = 3\), \(r_6 = 5\) and \(r_7 = 50\) and varied the time \(t\) over a range of \(t = 0.001\) to \(t = 0.1\). The result of our test is shown in Figure 1, where straight lines represent the rates used in the simulation while symbols show the rates as estimated by both methods. Note that for each data point in Figure 1 the consensus sequences used in the simulation had a different (randomly assigned) base composition. We also computed 95% confidence intervals by using 10 independent replications for each data point; these intervals are only shown for the rate estimates in which we see a significant deviation from the predefined values.

The two single nucleotide transitions \(r_5\) and \(r_6\) are accurately reproduced by both methods. For the CpG-specific rate, however, only M2 could reproduce the expected frequencies over the entire range, while M1 produced a slight but significant bias for \(t > 0.03\). Interestingly, there is a specific time range where M1 seems to be susceptible to biases resulting from the summation over the flanking neighbors in the objective function without correcting for the truncation error. To explore whether the summation over the flanking bases or the truncation error is primarily responsible for the bias produced by M1, we simulated an ensemble of trinucleotides using the same parameters as stated above. Since this dataset is already restricted to a trinucleotide scale, a truncation error cannot arise. This allows us to study in isolation the effect of the summation over the flanking neighbors. We observe that a similar bias (shown in Fig. 1) occurs when applying M1 to the data, which vanishes again if M2 is used instead (data not shown). We conclude from this that the bias produced by M1 must be primarily due to the summation over the flanking neighbors.

Returning to the simulation results shown in Figure 1, it is worth noting that the estimation error of the CpG-related rate was about an order of magnitude lower for M2 than for M1. Additionally, the importance of the boundary effect can also be seen by calculating the difference between \(r_7\) and \(r_6\) for each data point; these intervals are only shown for the rate estimates in which we see a significant deviation from the predefined values. The result \(\frac{1 - r_7}{r_7}\) averaged over the observed range. For \(r_7\) this yields 72%, but only 6% for \(r_5\) and <1% for the average transversion frequency.

### 2.4 Materials and statistics

Genome build hg18 (human) and mm8 (mouse) was downloaded from the UCSC browser (Kent et al., 2002). Human and mouse repeat information was extracted by RepeatMasker v. 3.1.2 and RM database version 20051025. All regression analyses were performed in R (http://www.r-project.org). In order to test for differences in the slope of two linear relations, a bootstrap test was implemented. Here, the residues from the regression rather than the data itself are resampled with replacement, as described in Davison and Hinkley (1997). To achieve a higher convergence rate of the bootstrap procedure and a more powerful test (Hall and Wilson, 1991; Peifer et al., 2005), studentized bootstrap was used as well. For each test 10,000 bootstrap samples were chosen.

### 3 RESULTS

First we compare and test the two context-dependent models discussed in Section 2. Since M1 is the basis for the results of Arndt et al. (2003a), by using that method we should be able to reproduce the main result of that work; the sudden decrease in the CpG-related frequency \(r_7\) after a given point in the past. To this end, we used M1 to analyze 494 repeat families, including all 38 families used in the study of Arndt et al. (Arndt and Hwa, 2004; Arndt et al., 2003a). This allows us to calculate the context-dependent genome-wide transition frequencies \(r_{5a}\), \(r_{6a}\) and \(r_{7a}\), and the average transversion frequency \(r_{7a} = \frac{(r_1 + r_2 + r_3 + r_4)}{4}\). In Figure 2 we reproduce Figure 4 from Arndt et al. (2003a), marking the 38 families used by that study with crosses. As expected, the 38 data points used by Arndt et al. (2003a) clearly reproduce the figure of the previous study, including the sudden change in the number of CpG-related transitions at \(r_{7a} \approx 0.025\). For the transition frequencies \(r_{5a}\) and \(r_{6a}\), a linear relationship becomes apparent. The slope of the straight lines yields: \(r_5/r_{7a} = 2.73 \pm 0.04\) and \(r_6/r_{7a} = 5.35 \pm 0.07\)—results compatible with the Arndt study. Including the other 456 repeat families confirms the decrease of \(r_{7a}\) after \(r_{7a} \approx 0.025\). While \(r_5/r_{7a} = 3.00 \pm 0.03\) changes slightly...
linear fit of $r_{tr}t$ to $CpG$ included in the boundary-sensitive M1. We repeated the same analysis for the mouse genome; results are methods to the same input data, the simulation study suggests that the bias. Indeed, a different result is found when we apply M2 of 326.3 for the model allowing a change of the data (having the lower BIC score of 324.6, compared to a score a constant BIC (Schwarz, 1978). It turns out that the simple model for which $r_{tr}t$ for truncation effects. In addition, the huge difference between of 326.3 for the model allowing a change of the $r_{tr}t$ rate is assumed is indeed sufficient to explain the $r_{tr}t$ no longer undergoes any significant change at all. This again illustrates the importance of correcting the data points of Figure 3 and 4 to a common time axis, we would have to expand the values on the $x$-axis of Figure 4 relative to those of Figure 3 by roughly a factor of two (up to the mouse–human speciation). Hence any point on the $x$-axis which is the same in both Figure 3 and 4 must map to significantly different points on this common time axis. If the predicted, sudden change in the $CpG$ substitution rates for human and mouse occurs at roughly the same $x$-axis point in both figures, it follows that these changes must map to two different points on the temporal axis. Therefore, the two shifts cannot both have happened at the time of the mammalian radiation. A more plausible explanation is that the sudden change is an artifact of the method of calculation.

Next we look at a human–mouse comparison of $r_5/r_{tr}$, $r_6/r_{tr}$ and $r_7/r_{tr}$ obtained by using M2. For $r_5/r_{tr}$, we observe a slight upward change to $r_5/r_{tr} = 3.14 \pm 0.02 \times 10^{-4}$, whereas for the other two rates we obtain $r_6/r_{tr} = 4.69 \pm 0.03$ and $r_7/r_{tr} = 42.3 \pm 0.5$ for the mouse, which is significantly smaller ($P < 10^{-4}$) than the corresponding values obtained for the human genome. The factors potentially responsible for this change are investigated in the discussion.

The most direct way to check the performance of a fitting procedure is to compare the fitted results to the original data. This is done in Figure 5 on the basis of the transition probabilities towards $CpG$ sites. Aligning the ancestral consensus sequence to the individual repeat copies, one can empirically compute the transition probability from a $CpG$ site (or non-$CpG$ site) on the consensus sequence to a $CpG$ site on today’s genome; thus computing the

Fig. 3. Substitution frequencies as in Figure 2, computed now with two alternative methods. Method 1 (open symbols) has been used in Arndt et al. (2003a) and consists in a fitting procedure optimizing a contracted log-likelihood function, see Equation (1), while method 2 (closed symbols) corrects for truncation errors arising by the restriction to trinucleotides and uses the objective function given in Equation (2).

Fig. 4. Here we have shown the same comparison of method 1 against method 2 as in Figure 3, fitted to mouse genomic data. Again, applying method 1, a shift of the $CpG$-related rate becomes apparent (open symbols), whereas this shift disappears when method 2 is applied (closed symbols). Interestingly, the sudden change in $r_{tr}t$ of the M1-based estimates is roughly at the same average transversion frequency as for the human genome.
Evidence against a change of the CpG-rate

Fig. 5. Probability of observing a CpG→CpG (left) and a non-CpG→CpG (right) state transition, versus the averaged transversion frequency representing the age of each repeat family. Upper two panels: data from the human genome; lower two panels: synthetic data from simulation. Original data (plus symbols) are computed from the alignment of the consensus sequences to modern repeat copies. Model predictions are based on the rates estimated either with method M1 (blue open square symbols) or with method M2 (filled red circles).

In Figure 2 beyond which the sudden change in CpG deviation similar to that already found in Figure 1. better agreement for M1, even though a bump is to be observed—a synthetic data. The lower two panels in Figure 5 indeed show a data than that of real data, M1 is expected to perform better using estimated for each repeat family α the age of each repeat family. Upper two panels: data from the human genome; lower two panels: synthetic data from simulation. Original data (plus symbols) can be inserted into the corresponding rate matrices to compute transitional probabilities from the models. Comparison of these model predictions to the actual transition probabilities allows us to judge the success of the respective fitting procedure. Figure 5 also shows the same comparison using the synthetic data from the simulation presented in Figure 1.

We observe that both the fast decay of the CpG sites and the build-up of CpG sites from non-CpG sites are well reproduced by M2, while method M1 fails to produce a satisfactory agreement for all data points beyond \( r_{tr}a_g = 0.025 \). Note that this is just the point in Figure 2 beyond which the sudden change in CpG occurs. Since the statistical noise is orders of magnitude lower for the simulated data than that of real data, M1 is expected to perform better using synthetic data. The lower two panels in Figure 5 indeed show a better agreement for M1, even though a bump is to be observed—a deviation similar to that already found in Figure 1.

4 DISCUSSION AND CONCLUSION

In this work, we reconsidered the context-dependent substitution rate model proposed by Arndt and colleagues (Arndt and Hwa, 2004; Arndt et al., 2003a). These studies have concluded that the CpG-specific substitution rate \( r_7 \) along the mammalian lineage is not constant in time, but underwent a drastic change about 90 Mya (corresponding to an average transversion frequency of \( r_{tr}a_g = 0.025 \), or at roughly the time of the mammalian radiation. Our main finding is that this result cannot be confirmed when we use a more precise objective function and correct for the trinucleotide truncation effect. When applying our modified model, the evidence for the asserted change of the CpG-specific rate after \( r_{tr}a_g = 0.025 \) vanishes; we see that the CpG-specific rate has been constant over time, is 48.3 times larger than the average transversion rate for the human genome, and 42.3 times larger than \( r_7 \) for the mouse genome.

Our analysis suggests that the sudden shift of the CpG-specific rate proposed by Arndt and colleagues (Arndt and Hwa, 2004; Arndt et al., 2003a) is an artifact of their method. We have a number of arguments supporting this claim. First, the method used by Arndt et al. leads to an inconsistency which comes to light if one compares rates based on the analysis of the mouse and the human genome. A sudden CpG rate shift associated with an event in the past, such as the mammalian radiation, would imply that the shift occurs at a single point in physical time—and thus the estimation should be independent of the organism from which it is derived. We have applied the method used by Arndt et al. to estimate frequencies \( r_{tr}a_g \) for both the human and the mouse genome (Figs 3, 4). This comparison reveals that the predicted CpG-rate shift occurs at a point in the past corresponding to the same transversion frequency for the mouse and the human genome. When we account for the different rates of substitution experienced by the two genomes, we realize that

\[
\begin{align*}
\text{CpG-content restricted to locations where either the consensus sequence has a CpG site or not. This probability is plotted against the averaged transversion frequency, which can be considered as a time axis. Again, each data point (plus symbols) corresponds to a single repeat family whose ‘age’ is given by the value on the x-axis. However, the rates } & \text{ are computed from the alignment of the consensus sequences to modern repeat copies. Model predictions are based on the rates estimated either with method M1 or M2, can be inserted into the corresponding rate matrices to compute transitional probabilities from the models. Comparison of these model predictions to the actual transition probabilities allows us to judge the success of the respective fitting procedure. Figure 5 also shows the same comparison using the synthetic data from the simulation presented in Figure 1. }
\end{align*}
\]

\[
\begin{align*}
\text{We observe that both the fast decay of the CpG sites and the build-up of CpG sites from non-CpG sites are well reproduced by M2, while method M1 fails to produce a satisfactory agreement for all data points beyond } & \text{ Note that this is just the point in Figure 2 beyond which the sudden change in CpG occurs. Since the statistical noise is orders of magnitude lower for the simulated data than that of real data, M1 is expected to perform better using synthetic data. The lower two panels in Figure 5 indeed show a better agreement for M1, even though a bump is to be observed—a deviation similar to that already found in Figure 1. }
\end{align*}
\]

\[
\begin{align*}
\text{4 DISCUSSION AND CONCLUSION} \\
\text{In this work, we reconsidered the context-dependent substitution rate model proposed by Arndt and colleagues (Arndt and Hwa, 2004; Arndt et al., 2003a). These studies have concluded that the CpG-specific substitution rate } & \text{ along the mammalian lineage is not constant in time, but underwent a drastic change about 90 Mya (corresponding to an average transversion frequency of } \text{ at roughly the time of the mammalian radiation. Our main finding is that this result cannot be confirmed when we use a more precise objective function and correct for the trinucleotide truncation effect. When applying our modified model, the evidence for the asserted change of the CpG-specific rate after } \text{ vanishes; we see that the CpG-specific rate has been constant over time, is 48.3 times larger than the average transversion rate for the human genome, and 42.3 times larger than } & \text{ for the mouse genome. }
\end{align*}
\]

\[
\begin{align*}
\text{Our analysis suggests that the sudden shift of the CpG-specific rate proposed by Arndt and colleagues (Arndt and Hwa, 2004; Arndt et al., 2003a) is an artifact of their method. We have a number of arguments supporting this claim. First, the method used by Arndt et al. leads to an inconsistency which comes to light if one compares rates based on the analysis of the mouse and the human genome. A sudden CpG rate shift associated with an event in the past, such as the mammalian radiation, would imply that the shift occurs at a single point in physical time—and thus the estimation should be independent of the organism from which it is derived. We have applied the method used by Arndt et al. to estimate frequencies } & \text{ for both the human and the mouse genome (Figs 3, 4). This comparison reveals that the predicted CpG-rate shift occurs at a point in the past corresponding to the same transversion frequency for the mouse and the human genome. When we account for the different rates of substitution experienced by the two genomes, we realize that}
\end{align*}
\]

\[
\begin{align*}
\text{CpG-content restricted to locations where either the consensus sequence has a CpG site or not. This probability is plotted against the averaged transversion frequency, which can be considered as a time axis. Again, each data point (plus symbols) corresponds to a single repeat family whose ‘age’ is given by the value on the x-axis. However, the rates } & \text{ are computed from the alignment of the consensus sequences to modern repeat copies. Model predictions are based on the rates estimated either with method M1 or M2, can be inserted into the corresponding rate matrices to compute transitional probabilities from the models. Comparison of these model predictions to the actual transition probabilities allows us to judge the success of the respective fitting procedure. Figure 5 also shows the same comparison using the synthetic data from the simulation presented in Figure 1. }
\end{align*}
\]

\[
\begin{align*}
\text{We observe that both the fast decay of the CpG sites and the build-up of CpG sites from non-CpG sites are well reproduced by M2, while method M1 fails to produce a satisfactory agreement for all data points beyond } & \text{ Note that this is just the point in Figure 2 beyond which the sudden change in CpG occurs. Since the statistical noise is orders of magnitude lower for the simulated data than that of real data, M1 is expected to perform better using synthetic data. The lower two panels in Figure 5 indeed show a better agreement for M1, even though a bump is to be observed—a deviation similar to that already found in Figure 1. }
\end{align*}
\]

\[
\begin{align*}
\text{4 DISCUSSION AND CONCLUSION} \\
\text{In this work, we reconsidered the context-dependent substitution rate model proposed by Arndt and colleagues (Arndt and Hwa, 2004; Arndt et al., 2003a). These studies have concluded that the CpG-specific substitution rate } & \text{ along the mammalian lineage is not constant in time, but underwent a drastic change about 90 Mya (corresponding to an average transversion frequency of } \text{ at roughly the time of the mammalian radiation. Our main finding is that this result cannot be confirmed when we use a more precise objective function and correct for the trinucleotide truncation effect. When applying our modified model, the evidence for the asserted change of the CpG-specific rate after } \text{ vanishes; we see that the CpG-specific rate has been constant over time, is 48.3 times larger than the average transversion rate for the human genome, and 42.3 times larger than } & \text{ for the mouse genome. }
\end{align*}
\]

\[
\begin{align*}
\text{Our analysis suggests that the sudden shift of the CpG-specific rate proposed by Arndt and colleagues (Arndt and Hwa, 2004; Arndt et al., 2003a) is an artifact of their method. We have a number of arguments supporting this claim. First, the method used by Arndt et al. leads to an inconsistency which comes to light if one compares rates based on the analysis of the mouse and the human genome. A sudden CpG rate shift associated with an event in the past, such as the mammalian radiation, would imply that the shift occurs at a single point in physical time—and thus the estimation should be independent of the organism from which it is derived. We have applied the method used by Arndt et al. to estimate frequencies } & \text{ for both the human and the mouse genome (Figs 3, 4). This comparison reveals that the predicted CpG-rate shift occurs at a point in the past corresponding to the same transversion frequency for the mouse and the human genome. When we account for the different rates of substitution experienced by the two genomes, we realize that}
\end{align*}
\]
the time periods corresponding to a given transversion frequency are different for the two organisms, implying that the identified shift occurred at different times in each organism's history. This result is thus not compatible with the idea that the time-point of the sudden CpG-rate shift corresponds to a real event at a specific point on a physical time axis.

We also find evidence against the asserted CpG shift from a direct comparison of the results of the Arndt method against those of our method. Our method uses the objective function $L_2$ [Equation (2)] for the optimization procedure and applies it to the boundary-corrected model that accounts for truncation errors. Arndt and colleagues have taken the simple context-dependent model, which allowed them to apply the objective function $L_1$, Equation (1). The crucial difference is that $L_1$ is based on a summation over the two flanking bases in the triplet state—something that has not been done in $L_2$. More specifically, an optimization based on $L_1$ searches for an optimal solution in a space of lower dimensionality [using $N(i,j,k,\ldots)$ as compared to an optimization based on $L_2$ which uses the full $N(i,j,k,\ldots)$] to estimate the rates. Thus, while $L_1$ uses less information (in this case disregarding the information about a whole triplet of bases $(i,j,k)$, any method using $L_2$ exploits all available information and hence is a more general approach.

We find the Arndt method cannot cope with an unfavorable signal-to-noise ratio. Using simulation data with an exceptionally high noise (e.g. statistical errors caused by the finite size of the sample, alignment errors and errors in the reconstruction of the ancestor sequences of the repeats) differences in the single nucleotide composition arising from different initial states are increasingly masked by the noise as the sequence approaches its equilibrium. Essentially, it is these differences that are needed to estimate the CpG-specific rate $r_7t_q$.

A further possible source of error in our analysis is related to the ancestral CpG dinucleotide frequency for highly diverged repeat families. As Figure 5 reveals, when the average transversion rate $r_7t_q = 0.025$ we expect about 93% of all CpG sites having to change to non-CpG sites. This might hamper the identification of ancestral CpG dinucleotides, and the reconstructed consensus sequences for highly diverged repeats may contain less CpG sites than were present in the actual ancestor. Using simulations we examined whether a partial loss of ancestral CpG-sites in the reconstruction leads to a substantial bias in the identification of $r_7$. To this end, 25% of the initial CpG sites of the synthetic data were altered to either TpG or CpA (with equal probability). Applying M1 to the simulated data we find that the method estimates $r_7/r_{tq}$ at about 63% of its actual value. In contrast, our estimation method (M2) does not result in a significant estimation bias. Our method does not require a perfect reconstruction of the ancestral sequence CpG sites. We correctly estimate the CpG-specific rate even with a significant number of reconstruction errors because our method is based on conditional transition probabilities and takes the CpG content of today's sequence into account. Furthermore, the ability to take the modern CpG content into account helps in the estimation of the CpG rate even if the data are affected by additional noise coming from the omission of some ancestral CpG sites. See the Supplementary Materials for the supporting simulation results.

In considering the points discussed above, we did look for a signature of CpG depletion in the consensus sequences of highly diverged repeats by correlating the CpG content of the ancestral sequence to the average transversion frequency. No clear trend became apparent. In contrast, if the reconstructed consensus sequences had a strong systematic CpG bias, then one would expect the CpG content on the consensus sequences would decay with the average transversion frequency.

Taken together, these results strongly suggest that the shift of $r_7t_q$ observed in Arndt et al. (2003a) is simply an artifact produced by the method. We note that the same model and method has been used in a number of studies (e.g. Arndt and Hwa, 2004; Arndt et al., 2003a, b; Ebersberger and Meyer, 2005; Meunier and Duret, 2004; Webster et al., 2006). The results obtained in these papers should be reconsidered in the light of the results presented here.

Our finding that the purported CpG-related substitution rate shift is a computational artifact is supported by a number of previous biologically oriented studies. Arndt et al. explained their shift with a sudden change in the methylation pattern, but to date there has been no report of a biological event which would explain such a shift. In contrast, such events have been identified at an earlier point in genomic history. A distinct change of the methylation pattern is known to have taken place only at the invertebrate–vertebrate boundary, roughly 450 Mya. Evidence for this change is presented in Tweedie et al. (1997). There it is observed that the methylation pattern is relatively constant over all vertebrate genomes (showing a high degree of methylation throughout),
whereas CpG sites in invertebrate genomes are predominantly non-methylated. This change of the methylation pattern can be attributed to a distinct change in the methyl-CpG binding proteins, as discussed in Hendrich and Tweedie (2003). In specific, the MBD2/3 protein is encoded by a single gene within invertebrate genomes, whereas the two genes MBD2 and MBD3 are only present in vertebrate genomes. It is likely that MBD2/3 represents the original methyl-CpG binding protein, that MBD2 and MBD3 are direct descendants of MBD2/3, and that this diversification plays an important role in the change of the DNA methylation pattern at the vertebrate–invertebrate boundary. Further studies have shown that this change of the DNA methylation pattern at the invertebrate–vertebrate boundary is also reflected by the sequence content between vertebrates and invertebrates. Using an odds ratio measure between the CpG dinucleotide content and the product of the single nucleotide content of C and G, the lowest values were obtained for vertebrates—indicating a high degree of CpG depletion (Karlin and Mrázek, 1997). Within the analyzed group of vertebrates all derived odds ratios are comparable and significantly lower than those of the invertebrates. In Cardon et al. (1994) the same odds ratio measure was computed for several mitochondrial genomes, showing that there is no significant change at the vertebrate–invertebrate boundary. Moreover, the odds ratios for the mitochondrial genomes are quite close to those of the nuclear genomes of the invertebrates. As no methylase is known to be active in the mitochondria, we might conclude that the observed change between vertebrates and invertebrates is due to a change in the methylation pattern. In addition, based on a simple context-dependent substitution model and assuming constancy of the substitution rates, Cooper and Krawczak (1989) estimated that the massive CpG depletion started roughly 450 Mya, thus at the vertebrate–invertebrate boundary. This result has been disputed in Jabbari et al. (1997), where a slight acceleration in depletion of CpG sites was observed for mammals and birds (amniotes) when compared to fish and amphibians (anamniotes). The authors explain this result by a change of body temperature resulting in a higher deamination rate for warm-blooded vertebrates. But neither a change at the anamniote–amniote boundary (about 320 Mya) nor the massive change of the methylation pattern at the vertebrate–invertebrate boundary (about 450 Mya) would explain a change in the methylation pattern 90 Mya, as claimed by Arndt et al. to explain the CpG-rate shift. Without evidence of such an event, our prediction of a constant CpG rate over the studied time span is the biologically more plausible result.

Hwang and Green (2004) suggest an alternative interpretation of the CpG-rate change at the time of the mammalian radiation. These authors find that CpG substitutions are relatively clock-like [in agreement with Kim et al. (2006)], while the single nucleotide substitutions are not consistent with the molecular clock hypothesis (see also Steiper et al., 2004; Yi et al., 2002). Because the single nucleotide substitutions, as opposed to the CpG substitutions, are prone to replication-dependent errors, it is conjectured that the CpG rate has remained constant whereas the rate of the other state transitions have decreased over time due to these replication-dependent errors. The combined effect of these findings would lead to a change of the CpG rate when plotted against the average transversion rate. In fact, if we compare the values of $rT/rF$ between human and mouse we find significantly smaller values for the mouse than for the human. This decrease is consistent with the relatively clock-like behavior of $rT/rF$, as discussed in Hwang and Green, (2004) and Kim et al. (2006), because the increase of the average transversion frequency is larger than the increase in CpG-rate when passing from human to mouse. Consequently, the ratio $rT/rF$ is expected to be smaller for the mouse. Conversely, $rT/rF$ should have similar values for human and mouse when estimated from repeats which are older than the speciation time between these species. Therefore, a small upward shift should be visible for the ancient repeats of the mouse lineage. But since the noise on the data is so high, such a shift cannot be reliably examined with the proposed method. Thus, further studies are needed to completely clarify fine structure of the time dependency of the CpG rate.

ACKNOWLEDGEMENTS

The authors would like to thank Philip Green, Ross Harrison, Webb Miller and Svitlana Tyekucheva for their help with this research, Laura Tabacca and David Karro for their help in writing this article, and Richard Burhans and Nathan Corora for their technical support.

Funding: National Institutes of Health (SK01HG003315 to J.K.); Science Foundation (FWF) (P18762 to M.P.).

Conflict of Interest: none declared.

REFERENCES


