Variations in Substitution Rate in Human and Mouse Genomes

H. H. von Grünberg,[1] M. Peifer,[2] J. Timmer,[2] and M. Kollmann[3,2,*]

[1]*Universität Graz, Institut für physikalische Chemie, 8010 Graz, Austria*
[2]*Universität Freiburg, Institut für Physik, 79104 Freiburg, Germany*
[3]*Università di Palermo, Dip. di Fisica e Tech. Relative, Palermo, Italy*

(Received 18 December 2003; published 11 November 2004)

We present a method to quantify spatial fluctuations of the substitution rate on different length scales throughout genomes of eukaryotes. The fluctuations on large length scales are found to be predominantly a consequence of a coarse-graining effect of fluctuations on shorter length scales. This is verified for both the mouse and the human genome. We also found that the relative standard deviation of fluctuations in substitution rate is about a factor three smaller in mouse than in human. The method allows furthermore to determine *time-resolved* substitution rate maps of the genomes, where the corresponding autocorrelation functions quantify the velocity of spatial chromosomal reorganization.

DOI: 10.1103/PhysRevLett.93.208102 PACS numbers: 87.23.Kg

The detailed knowledge of the mechanisms of mutations in living organisms is of fundamental importance for understanding the evolution of genomes. The basis of development in multicellular organisms is the cell devision cycle and mutations are changes made in DNA during the process of chromosome replication. Nonlethal mutations occurring in the germline are passed onto the next generation. A large fraction of these mutations are substitutions of single nucleotides (A,T,G or C) while other mutations are due to insertions or deletions of one or more nucleotides into the DNA sequence. In recent time there has been growing acceptance that the substitution rate is not spatially constant inside the genomes of mammals [1–8]. This is a surprising result at first sight, as nucleotide substitutions resulting from replication errors should be fairly independent on the actual position, at least on large length scales. However, this picture is too simple, as there exist strong correlations in the mammalian genome between mutation rate and other evolutionary rates (e.g. recombination rate) [8–10]. Although the origin of the substitution rate bias in genomes of mammals is unknown, it is highly important to quantify the length and time scales where changes are occurring. This is because the amount of conserved sequences between the genomes of different species depends crucially on the fluctuations of the local mutation rate. However, these conserved sequences give the best insight into which parts of the genomes carry function and are therefore under selective constraint.

Here, we present a method to calculate substitution rates in genomes of eukaryotes. The basis of our method is to make use of interspersed repeats [8,11,12]. Interspersed repeats are sequences of $3 \times 10^2 - 5 \times 10^3$ base pairs in length whose copy was inserted up to 10^6 times in the human genome. Each copying machinery has only worked for a short time in the history of the organism and from that time on, the copies of a specific repeat type have accumulated substitutions and other mutations. Because of the large number of copies the original sequence can be reconstructed quite accurately. Differences between a given copy (repetitive element) of an interspersed repeat and its consensus sequence allows us to estimate the mutation rate at the position of this repetitive element if the time is known when the copying machinery was active (c.f. Refs. [11]). In the human genome there have been classified more than 600 different interspersed repeats which occupy in total more than 45% of the genome. This gives us a large set of sequences at hand which is most likely under no selective constraint [12]. We visualize interspersed repeats as "measurement devices" for the underlying local substitution rate. This requires us to solve three major problems: (i) single repetitive elements are usually too short and have on average accumulated too few substitutions to give a reliable estimate for the substitution rate at their position in the genome; (ii) the elements show a very broad length distribution (implying that our "measurement devices" differ in their "sensitivity", which is proportional to their length); and (iii) the repetitive elements belonging to different types differ in general in their age (thus, the measurement devices have been measuring over different periods of time). Now the major theoretical task is to derive from the large amount of repetitive elements, which have large diversity in their age and sensitivity, reliable information about the underlying substitution rate at different positions and at different times in the genome.

For our analysis we use a set of 649 different types of interspersed repeats [13] as identified by the REPEATMASKER program [14] in the complete genomes for human (hg16) and mouse (mm4) from the UCSC bioinformatics site. The consecutive sequences are taken from the RepBase database [11]. To investigate variations in the substitution rate on different length scales we divide the two genomes into equally sized partitions. The total number of bases of all repeats of type α in the partition γ is denoted by $N_{\alpha\gamma}$ and the corresponding number of base changes from the consensus sequence $i \rightarrow j$ with $i, j \in \{\mathrm{A, T, G, C}\}$ is given by $k_{\alpha\gamma}^{ji}$. Next, we define for each partition (γ) and each repeat type (α) the average divergence, $Q^{ji}_{\alpha\gamma}$, which is the probability to observe a substitution of base i in the consensus sequence by base j in the corresponding genomic sequence. For statistically independent substitutions, the probability to find $k^{ji}_{\alpha\gamma}$ mismatches in interspersed repeats of type α located in the partition γ, is given by

$$P(\mathbf{k}_{\alpha\gamma}|\mathbf{Q}_{\alpha\gamma}, \mathbf{N}_{\alpha\gamma}) = \prod_i N^i_{\alpha\gamma}! \prod_j \frac{Q^{ji^{k^{ji}_{\alpha\gamma}}}_{\alpha\gamma}}{k^{ji}_{\alpha\gamma}!} \qquad (1)$$

with $N^i_{\alpha\gamma} = \sum_j k^{ji}_{\alpha\gamma}$ and $N_{\alpha\gamma} = \sum_i N^i_{\alpha\gamma}$ where the products run over the set $\{\mathrm{A, T, G, C}\}$. For sufficiently large partitions, P converges to a Gaussian distribution with mean $\langle k^{ji}_{\alpha\gamma}\rangle = N^i_{\alpha\gamma} Q^{ji}_{\alpha\gamma}$ and variance $\langle\langle k^{ji}_{\alpha\gamma}\rangle\rangle = N^i_{\alpha\gamma} Q^{ji}_{\alpha\gamma}(1 - Q^{ij}_{\alpha\gamma})$. Thus, the average divergence, $\mathbf{Q}_{\alpha\gamma}$, can be estimated from the knowledge of $\mathbf{k}_{\alpha\gamma}$ and $\mathbf{N}_{\alpha\gamma}$ by a least mean square fit.

To gain information from the quantities $\mathbf{Q}_{\alpha\gamma}$ about the local substitution rate in partition γ, denoted by $m_\gamma(t)$, we now introduce a microscopic model for base substitutions. Statistically independent changes of a base at a given position in the genome at time t can be described by the following Master equation

$$\partial_t \mathbf{p}(t) = \mathbf{w}(t) \cdot \mathbf{p}(t), \qquad (2)$$

where the elements $p_b(t)$ of the vector $\mathbf{p}(t)$, with $b \in \{\mathrm{A, T, G, C}\}$, denote the probability of observing the base b at a certain site in the genome at time t given the initial state $\mathbf{p}(-t_0)$. The transition matrix $\mathbf{w}(t)$ has the elements $w_{bb'}(t) = m_\gamma(t) q_{bb'}(t)$ for $b \neq b'$ and $w_{bb}(t) = -\sum_{b' \neq b} m_\gamma(t) q_{b'b}(t)$ [15]. Here, the elements $q_{bb'}(t)$ of the matrix $\mathbf{q}(t)$ are the transition probabilities that a base b' mutates to a base b whenever a substitution occurs.

From the relation $\mathbf{p}(0) = \mathbf{Q}_{\alpha\gamma} \cdot \mathbf{p}(-t_\alpha)$ follows that the average divergence from the consensus sequence is given by

$$\mathbf{Q}_{\alpha\gamma} = \exp\left[\int_{-t_\alpha}^{0} m_\gamma(t) \mathbf{q}(t) dt\right]. \qquad (3)$$

Here, t_α denotes the time distance from today to the moment when the interspersed repeat of type α was inserted into the genome for the first time. We emphasize that nearest neighbor effects have an impact on the substitution rate pattern (c.f. Ref. [16]) but seem not to dominate the fluctuations in substitution rate on the large length scales considered [8]. Now, we define the genome-wide average substitution rate as $m(t) = 1/Z \sum_{\gamma=1}^{Z} m_\gamma(t)$, with Z the number of partitions, and the spatial deviations from this rate as $\tau_\gamma(t) = m_\gamma(t) - m(t)$. We further introduce the time-averaged mean substitution rate by $\bar{m}_\alpha = 1/t_\alpha \int_{-t_\alpha}^{0} m(t) dt$ and the corresponding deviations as $\bar{\tau}_{\alpha\gamma} = 1/t_\alpha \int_{-t_\alpha}^{0} \tau_\gamma(t) dt$. For our purposes we can take to good approximation the transition matrix as time-independent $\mathbf{q} = \mathbf{q}(t)$. With these definitions, the argument in the exponential of Eq. (3) reads $\mathbf{q} \int_{-t_\alpha}^{0} m_\gamma(t) dt = \mathbf{q}(\bar{m}_\alpha t_\alpha + \bar{\tau}_{\alpha\gamma} t_\alpha)$. It is clear now that we can obtain only the *time-averaged* quantities $\bar{m}_\alpha$, $\bar{\tau}_{\alpha\gamma}$ from the knowledge of $\mathbf{Q}_{\alpha\gamma}$.

The repetitive elements can be divided into lineage-specific repeats (defined as those introduced by transposition after the divergence of human and mouse) and ancestral repeats (defined as those already present in the common ancestor). In the following we first analyze ancestral repeats, as these were used in previous work to report on local variations in substitution rates [8,10]. By taking ancient repeats from a sufficiently narrow time window close to the speciation time (ca. 70 Myr) we can make the approximations $\bar{\tau}_\gamma \approx \bar{\tau}_{\gamma\alpha}$ and $\bar{m} \approx \bar{m}_\alpha$. The time scales are given by setting the substitution rate of the transversions $q_{\mathrm{AC}} = q_{\mathrm{TG}} = 0.05$ to a fixed value and defining $\bar{m}_{\mathrm{Human}} = 1$ and $\bar{m}_{\mathrm{Mouse}} = 1$. One should note, however, that by using physical time scales the mean substitution rate in the mouse lineage is about a factor two larger than in the human lineage since the time of speciation (c.f. Ref. [8]).

We start our calculations by determining iteratively the transition matrix, $\mathbf{q}$, and the individual age of the repeats, t_α, by maximizing the corresponding *log-likelihood* functional of Eq. (1) using a multiple shooting method [17]. The multiple shooting method integrates Eq. (2) and performs a least mean square fit to determine the free parameters in a robust way. As start values for the iteration scheme we use equal transition rates (Jukes-Cantor approximation), $q_{bb'} = 1/4$ for $b \neq b'$, resulting in $t_\alpha = -1/(\bar{m}) \ln[1 - (4/3) \sum_{\gamma=0}^{Z} k_{\alpha\gamma} / \sum_{\gamma=0}^{Z} N_{\alpha\gamma}]$ with $k_{\alpha\gamma} = \sum_{i \neq j} k^{ji}_{\alpha\gamma}$. The resulting transition rates are given after convergence by $q_{\mathrm{AC}} = q_{\mathrm{TG}} = 0.050$, $q_{\mathrm{AT}} = q_{\mathrm{TA}} = 0.056$, $q_{\mathrm{CG}} = q_{\mathrm{GC}} = 0.057$, $q_{\mathrm{CA}} = q_{\mathrm{GT}} = 0.070$, $q_{\mathrm{AG}} = q_{\mathrm{TC}} = 0.172$, $q_{\mathrm{GA}} = q_{\mathrm{CT}} = 0.372$, and are thus close to the one reported in Refs. [12,18]. After determination of $\mathbf{q}$ and t_α from a whole genome analysis we partition the genomes to estimate the time-averaged local substitution rates, $\bar{\tau}_\gamma$, using again maximum likelihood estimation. The results for ancient repeats within the human genome are shown in Fig. 1 for a partition size of 10^6 base pairs (Mbp) using the time window [0.479, 0.675]. The speciation time in our time scales is given by ≈ 0.49. In the inset of Fig. 1 we show results for the local substitution rates for chromosome 22. The local substitution rates are averaged over five neighboring partitions of 10^6 base pairs in steps of 10^6 base pairs. The differences in the local substitution rates as measured by long interspersed nucleotide elements (LINEs) (see Ref. [12] for informations about LINEs) and measured by the whole set of annotated repeats is mostly due to still significant variations of the substitution rate on length scales smaller than 10^6 base pairs. As LINE repeats comprise just one third of the total amount of repetitive elements within this time window

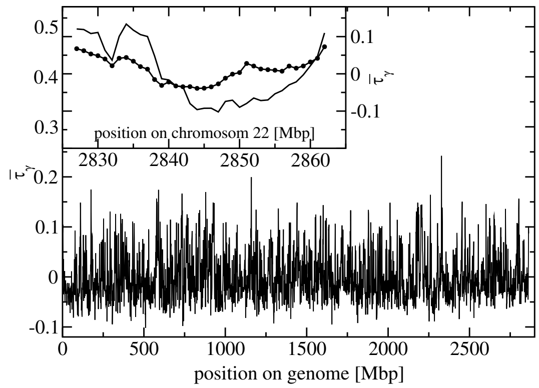

FIG. 1. Time-averaged local substitution rate of the human genome for 10^6 base pairs partition size using ancestral repeats. The chromosomes are concatenated in increasing order. The inset shows the results for chromosome 22. The repetitive elements of only the LINE class (solid line) is compared with all major classes (LINE, SINE, LTR, DNA) of repetitive elements (solid line with circles).

(more than 6×10^4 base pairs per partition) the sampling error is clearly larger than in the full set of repetitive elements. The standard deviations of these fluctuations, $\bar{\sigma}_A = (1/Z\sum_{\gamma=1}^{Z}\bar{\tau}_\gamma^2)^{1/2}$, are shown in Fig. 2 for different partition sizes. On length scales larger than 5×10^6 base pairs the variations in substitution rates result from coarse graining of statistical independent partitions of smaller size as can be shown by rescaling the standard deviation, $\bar{\sigma}_A$, with the number of partitions, Z, (c.f. inset of Fig. 2). If the fluctuations were statistically independent for the highest spatial resolution, $Z = Z_{\max} = 2864$ (10^6 base pairs partition size), then $Z^{-1/2}\bar{\sigma}_A$ would be constant for all $Z < Z_{\max}$ and this seems to be the case for partitions of size $> 5 \times 10^6$ base pairs in the human genome. By comparing the variations in substitution rate between mouse and human we find an almost three fold higher standard deviation, $\bar{\sigma}_A$, for the human than for the mouse lineage, Fig. 2. The reason can be seen by a much higher rate of deletions and insertions in mouse resulting in a higher rate of chromosome reorganizations in mouse than in human. Consequently, we expect the correlations among adjacent partitions to be smaller in mouse and indeed this is found by our analysis as shown in the inset of Fig. 2. Here, the rescaled standard deviation of fluctuation remains almost constant for the mouse lineage even for partition sizes below 5×10^6 base pairs. We also checked the bias due to the incomplete coverage of the genomes by repetitive elements by performing block wise bootstrap [19]. Confidence intervals of 95% for determining the values $\bar{\sigma}_A$ are shown in Fig. 2 by the error bars.

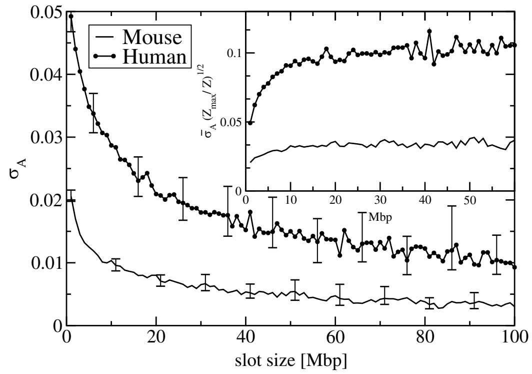

FIG. 2. Standard deviation $\bar{\sigma}_A$ of the local fluctuations of the substitution rate versus the size of the partitions for the human and mouse genome. The inset shows the normalized standard deviations $\bar{\sigma}_A/\sqrt{Z}$.

So far, we have computed $\bar{\tau}_\gamma$ for a specific time window for the class of youngest ancient repeats. Including also the lineage-specific repeats, we can repeat our optimization procedure but now using all repetitive sequences but grouped in eight equally distant age classes within the time interval [0.17, 0.77]. This time interval is obtained from the ages of all annotated repeat types in RepBase by removing the 5% youngest and oldest repeats. We then approximate $\bar{\tau}_\gamma^{(i)} = \int_{-(t_i+t_{i+1})/2}^{0}\tau_\gamma(t)dt \approx 1/M_i\sum_\alpha\bar{\tau}_{\alpha\gamma}$ where the sum runs over the M_i interspersed repeats, α, whose ages are within time window $i \in \{1, .., 8\}$, $t_i < t_\alpha < t_{i+1}$. We recall that $\bar{\tau}_\gamma^{(i)}$ is a time-averaged quantity, so it does not reflect the strength of fluctuations of the substitution rate $\tau_\gamma(0)$ as found today in the human and mouse genomes. The biological reason for $\tau_\gamma(t)$ for $t < 0$ being different from $\tau_\gamma(0)$ is because by reorganizations of the chromosomes, i.e., by insertions and deletions of sequences, the local substitution rate as measured by repetitive elements changes in time. This effect can be due to changes of base compositions (e.g., GC content) or simply by a shift of repetitive elements by insertions and deletion in regions of different local substitution rate. It is then clear that the substitution rate fluctuations for ancestral repeats as shown in Fig. 1 are significantly smaller in amplitude than the true (i.e., actual) variations in substitution rate $\tau_\gamma(0)$. To give a good estimate of the magnitude of the true fluctuations we have to include the effect of genome reorganizations in our model.

Within a certain large partition γ of size >5 Mbp the number of repeats belonging to the same time window is sufficiently large and the substitution rate across this partition is the result of coarse graining over almost independent fluctuations on smaller length scale, c.f. inset of Fig. 2. Therefore, we can employ the central limit theorem to predict that the distribution of $\tau_\gamma(t)$ will be close to a Gaussian distribution. This gets supported by our analysis of ancient repeats, c.f. Fig. 1. Furthermore, as the time scales for genomic rearrangements to become fixed within a population are much smaller than the relevant time scales of our model the underlying process which changes the local substitution rate can be taken to be Markovian. Then, assuming this process to be quasistationary, one is tempted to write for the actual variation of the local mutation rate, $\tau_\gamma(t)$, within a continuous time

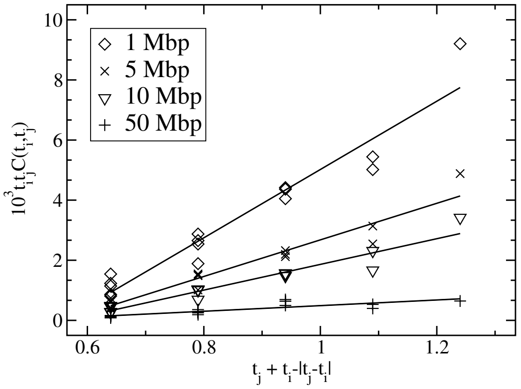

FIG. 3. The time-averaged correlation function $t_i t_j C(t_i, t_j)$ versus $t_i + t_j - |t_i - t_j|$ using eight equally sized time windows within the interval [0.17, 0.77]. The lines are determined by linear regression of Eq. (5), which determines the parameters a and σ for different partition sizes. We have restricted the time correlations to values $|t_i - t_j| \geq 0.15$ and $t_i, t_j > 0.32$, leading to small contributions of the exponential functions in the last line of Eq. (5) and thus justifying a linear fit to the data.

model

$$\partial_t \tau_\gamma(t) = a(t)\tau_\gamma(t) + \eta_\gamma(t). \qquad (4)$$

This is because by Doob's theorem it is essentially the only process satisfying the conditions stated above. Here, $a(t)$ is slowly varying, reflecting changes in the genome-wide substitution rate and $\eta_\gamma(t)$ can be chosen to be white noise with zero mean on the time scales considered. The autocorrelation function of the process, Eq. (4), is exponentially decaying. Thus, we obtain for the autocorrelation function $C(t_i, t_j) = \frac{1}{Z}\sum_{\gamma=0}^{Z} \bar{\tau}_\gamma^{(i)} \bar{\tau}_\gamma^{(j)}$ the expression

$$\begin{aligned} C(t_i, t_j) &= \frac{1}{t_i t_j}\int_0^{t_i}\int_0^{t_j} \sigma^2 \exp[-a|t - t'|]dtdt' \\ &= \frac{\sigma^2}{a t_i t_j}\bigg[t_i + t_j - |t_i - t_j| \qquad (5) \\ &\quad + \frac{1}{a}(e^{-at_i} + e^{-at_j} - e^{-a|t_i - t_j|} - 1)\bigg]. \end{aligned}$$

with $\sigma^2 = 1/Z\sum_{\gamma=0}^{Z} \tau_\gamma(t)\tau_\gamma(t)$ and a decay rate $a = a(t)$. Taking the fit parameters σ and a time independent is a reasonable approximation for the human species (c.f. Refs. [8,12]). Figure 3 shows the correlation function, $tt'C(t, t')$, for the human lineage. The free parameters $\sigma = \sigma(Z)$ and a in Eq. (5) are obtained by a least mean square fit from the data. For the partition sizes (1, 5, 10, 50) in units of 10^6 base pairs, we obtain the values $a = (1.79, 1.79, 1.78, 2.09)$ and the corresponding standard deviations are given by $\sigma = (0.079, 0.058, 0.049, 0.023)$. Thus the standard deviation for the fluctuations in substitution rate as found today in the human genome are about a factor 1.7 larger as found in the analysis using ancestral repeats $\sigma_A = (0.050, 0.036, 0.028, 0.013)$ (c.f. Fig. 2). The physical time for the correlations to decay to e^{-1} of their maximum values is about 85 Myr and thus slightly larger than the speciation time of human and mouse (70 Myr).

This decay time gives us an impression how fast chromosomal reorganizations alter local substitution rates in the human genome. The results also justify *a posteriori* the assumption of quasistationarity for our model, Eq. (4).

A central result of this work is that correlations of spatial fluctuations in substitution rate occur only on length scales *smaller* than 5×10^6 base pairs for the human and mouse genome. Therefore, differences in mean substitution rates between chromosomes, as reported in Ref. [12], arise simply from a coarse-graining effect. Furthermore, a time-resolved analysis has shown a significant underestimation of these fluctuation when measured by ancestral repeats as done in previous work [8,10]. For the case that the decay time for variations in the substitution rate exceeds the speciation time for any two species, a higher amount of conserved sequences between these species can be expected.

This research has been supported by Marie Curie Followship of the Europ. Comm. Progr. improving the Human Research Potential under contract no. HPMD-CT2001-107 in collaboration with INFM and the DFG.

*Corresponding author
Electronic address: markus.kollmann@physik.uni-freiburg.de

[1] K. Wolfe *et al.*, Nature (London) **337**, 283 (1989).
[2] D. Casane *et al.*, J. Mol. Evol. **45**, 216 (1997).
[3] G. Matassi *et al.*, Curr. Biol. **9**, 786 (1999).
[4] M. Nachman and S. Crowell, Genetics **156**, 297 (2000).
[5] E. Williams and L. Hurst, Nature (London) **407**, 900 (2000).
[6] F. Chen and W. Li, Am. J. Hum. Genet. **68**, 444 (2001).
[7] M. Lercher *et al.*, Mol. Biol. Evol. **18**, 2032 (2001).
[8] R. Waterston *et al.*, Nature (London) **420**, 520 (2002).
[9] A. Eyre-Walker and L. Hurst, Nat. Rev. Mol. Cell Biol. **2**, 549 (2001).
[10] R. Hardison, Genome Res. **13**, 13 (2003).
[11] J. Jurka, Trends Genet. **16**, 418 (2000).
[12] E. Lander *et al.*, Nature (London) **409**, 860 (2001).
[13] We used all repeats annotated in 'Repbase9.05'' at http://www.girinst.org.
[14] A. F. A. Smit and P. Green, RepeatMasker at http://repeatmasker.org.
[15] W. Li and D. Graur, *Molecular Evolution* (Sinauer Associates, Massachusetts, 1991).
[16] P. F. Arndt, D. A. Petrov, and T. Hwa, Mol. Biol. Evol. **20**, 1887 2003.
[17] H. Voss *et al.*, in *Chaos and its Reconstruction* (Nova Science Publishers, Inc., New York,2003).
[18] The apparent differences are due to the inclusion of CpG sites in our model. The question whether or not to include CpG sites, in a model which ignores nearest neighbor effects, would be important if one aims at absolute values of the substitution rates.
[19] M. Peifer *et al.*, Biom. J. (to be published).