Evidence on DNA slippage step-length distribution

Branko Borštnik* and Danilo Pumpernik
National Institute of Chemistry, Hajdrihova 19, SI-1001 Ljubljana, Slovenia
(Received 18 June 2004; revised manuscript received 30 November 2004; published 28 March 2005)

A simple model based on a master equation is constructed in order to reveal the details of the mutational events modifying simple sequence repeats in the human genome, A database of simple repeats together with their flanking sequences comprising approximately 10^5 entries from all 24 human chromosomes was constructed. By aligning the pairs of fragments of sequences containing the repeat elements, the matrices that count the number of slippage events were evaluated. The counts were then used as a target to be reproduced by our theoretical model, in which the elongation and shortening of the repeats proceed through a mechanism in which the step lengths exhibit a decaying distribution in the form of an inverse power law rather than through one nucleotide extension or deletion, which was the most frequent supposition in previous studies.

DOI: 10.1103/PhysRevE.71.031913 PACS number(s): 87.16.Ac, 87.23.Kg, 87.10.+e, 82.39.−k

I. INTRODUCTION

The acquisition of the nucleotide sequences of entire genomes of several biological species, including the human genome, improves the insight into biological processes on the molecular level and enables one to construct physical models. At the moment the understanding of organismal functioning and development is still buried deeply among the fine details of genome organization. The availability of information about nucleotide sequences enables us to obtain partial answers. One of the challenges is posed by the very high content of repeat sequences, which is close to 50% in the human genome [1]. Repeats can be of tandem or interspersed character. Interspersed repeats are the products of transposon activities, while tandem repeats have their source in repeat amplification mechanisms. Transposons are special sequences that possess the ability to form new copies elsewhere in the genome. Not all of them can multiply themselves autonomously. Examples of autonomous elements are long interspersed repeat elements (LINE's) or retroviruslike elements, each of them comprising several thousands of nucleotides. Autonomous transposon elements mediate in the proliferation of other transposon elements such as short interspersed repeats (SINE's) a few hundreds of nucleotides in length, which exhibit the highest copy number (one and half million) in the human genome. Altogether the interspersed repeats represent more than 40% of the human genome. The most widespread SINE elements are so called *Alu* sequences whose elements comprise approximately 330 nucleotides. Studies of the internal structure of the interspersed repeats reveal the stretches of tandem repeats that are integrated in the LINE or SINE sequence itself, and also the flanks of interspersed repeats are populated by tandem repeats. Besides the location in the vicinity of LINE or SINE sequences, tandem repeats are also scattered throughout the chromosomes. Repeats in nucleotide sequences have been linked to the notion of molecular parasitism [2].

In this work we shall focus our attention on the category of tandem repeats with monomeric units of minimal complexity. Such repeats are usually called microsatellites or short simple repeats (SSR's). They appear predominantly in noncoding regions [3,4], but also in the regions where proteins are coded [5].

The SSR content of the human genome is approximately 2%. This is far above the expectation on the basis of the hypothesis of a random nucleotide sequence. One can hypothesize that an overrepresented class of sequences can exist only if a special mechanism of repeat amplification [6] exists. It turns out that such a mechanism is present in the process of either DNA synthesis or DNA recombination. The repeat amplifying mechanism has its roots in the fact that the translational symmetry of the template strand produces degenerate energetic states in the landscape of the template DNA strand–enzyme and nascent DNA strand–enzyme interactions. Such a degeneracy makes the mechanism of DNA synthesis vulnerable to errors in the form of so called slippage events that occur when the DNA ribbon slips out of order to a new position relative to the polymerase enzyme, forward or backward for one or several monomeric repeat units. If the template strand slips backward, for instance, the resulting nascent strand will be prolonged; in the opposite case shortening will result. A similar mechanism is operative also in the case of another important molecular cellular process: DNA recombination at meiosis. Also in this case the repetitive DNA sequence represents a drawback since a vital phase of recombination is the pairing of complementary strands of highly homologous regions of maternal and paternal regions of DNA sequences. In repetitive regions the process of homologous pairing has multiple realizations and with a certain probability the recombination process results in the prolongation of one and the shortening of the other newly recombined strand. The above mentioned mechanisms become operative only above a certain threshold repeat length that is approximately ten base pairs. The value of the threshold repeat length is approximately equal to the extent of the contact between the DNA helix and polymerase enzyme [7].

*Electronic address: branko@hp10.ki.si

Since we are building our strategy on the statistics of the events that can be inferred by comparing pairs of sequences having a common ancestor, it is important to note that there exists also another mechanism which contributes to the growth of DNA content. This is the segmental duplication mechanism which generates new copies of DNA segments elsewhere on the same chromosome, or on another chromosome. The lengths of the segments being copied are usually on the megabase scale. Hsieh and co-workers [8] have shown that segmental duplication is a fundamental ingredient of genome evolution. For our purpose the phenomenon of segmental duplication mechanism is extremely important because it provides us with an ample amount of repeats positioned in closely related flanking sequences, which we analyze with the purpose of getting insight into the SSR evolutionary dynamics.

Besides repeat elongations and repeat shortenings, point mutations also modify the repeats. The results of point mutation are two shorter repeats or, in the case that the mutation takes place too close to the left or to the right end of the repeat, depending on the criterion for the minimum repeat length, only one or none of the remaining segments may retain the repeat character. The combined effect of slippage and point mutational mechanisms can be modeled in various ways [9]. The models should be able to reproduce, as well as the repeat length distribution, also the positional correlations between the sites where the repeats are located along the DNA sequence. It was shown by Peng *et al.* [10] that nucleotide-nucleotide correlation in the DNA sequences can be described by an inverse power law, but it was also shown that in specific cases the correlation functions decay exponentially [11]. Holste *et al.* [12], who analyzed one of the human chromosomes, found that the nucleotide-nucleotide correlation functions exhibit a clear inverse power law decay, and they demonstrated that the distribution of interspersed repeats and SSR's plays a significant role in shaping the overall picture of nucleotide distributions.

In our previous work [13] the positional correlations between the repeats were used as a constraint together with the SSR length distribution in constructing a model that allowed the slippage process to possess asymmetry in the elongation versus shortening of the SSR. We demonstrated that the model reproduces satisfactorily the properties of SSR's in the human genome. An obvious shortcoming of our approach was due to the fact that the model allowed only slippage events with unit step length $\Delta n = \pm 1$. In this work we explore how far from reality this approximation is. By examining many pairs of repeats together with their environments we scrutinize the flanking sequences on both sides of the repeats to detect evidence about a common history. If the flanking regions of two repeats possess high enough homology, they are most probably descendants of a common ancestral sequence. Similar or homologous sequences are believed to be the result of a divergence, be it in the form of speciation (appearance of new biological species) or in the form of copying DNA segments from one site to another site of the same chromosome or from one chromosome to another one. To detect the mutations that occur after speciation one should compare two paralogous (having the same function) DNA sequences belonging to the members of two closely related species. We limit ourselves to the human species where one can perform so called orthologous interchromosomal and intrachromosomal comparisons that retrieve the homologies resulting from transposon activities or segmental duplications. Transposon activities in human species are declining during the last several millions of years and today the comparisons that track the transposing processes uncover low homologies with several overlapping events taking place at the same site. The segmental duplications, on the other hand, seem to occur with a rather uniform pace and the total amount of sequences that were duplicated and exhibit today a similarity greater than 90% is approximately 3.3%.

The subjects of our analyses are homologous pairs of human sequences containing repeats. When finding such a case one should try to determine whether the difference between the repeats themselves, if there is any, stems from a single repeat modification event or whether there were several modifications superimposed on the particular repeat. In what follows we shall work out a strategy that will try to reveal the details of the repeat elongation and shortening process.

The most frequent short simple repeats in human genome are mononucleotides (a/t). They are followed by (at/at), (ac/gt), and (ag/ct) repeat elements whose probabilities of appearance exhibit a decreasing order. There are approximately 2×10^6 nucleotides within polyadenine tracts longer than ten base pairs. All the other categories of SSRs are at least five times less abundant and therefore we shall base our work on analyses of (a/t) repeats. This choice is to some extent problematic and (a/t) repeats are usually left aside arguing that "polyadenine tracts are usually associated with *Alu* sequences and hence subject to special selection constraints" (a quotation taken from Ref. [14]). In order to reduce the objections against the (a/t) repeats we have omitted from our analyses the polyadenine tracts which are parts of *Alu* sequences.

II. ANALYSES OF HUMAN POLYADENINE REPEATS

The sequences of all 24 human chromosomes from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens were analyzed and several files of polyadenine elements together with their environments in the form of fragments of left and right flanking sequences were created. The lengths of the flanking fragments were set arbitrarily to 20 nucleotides at each side of the repeat. The poly adenine tracts were required to have a minimum length of 10 nucleotides. Further, we only took into consideration the repeats whose flanks possess high enough information content. This means that the repeat elements whose flanks exhibited repeat structure were discarded. The resulting database extracted from human genome nucleotide sequences encompassed more than 3×10^5 repeats.

The flanks of all the repeats were mutually compared (the left flank with the left flank and the right flank with the right flank). When the homology was above 90% the information about the number of such cases was stored into the replacement frequency matrices $\tilde{A}^{(h)}(m,n)$. The (m,n) pair of indices refers to the lengths of the two repeats whose flanks were

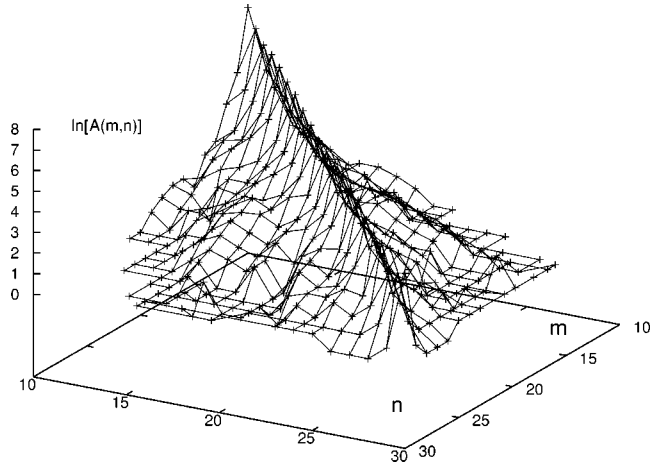

FIG. 1. Plot of the mutation frequency matrix $\tilde{\mathbf{A}}^{(h)}$ within the class $h=100\%$. The two horizontal axes refer to the lengths of the two polyadenine repeats that are compared and along the vertical axis the logarithm of the number of the pairs found for m,n pair of repeat lengths is plotted.

compared. The superscript h refers to the homologies between the flanks ($h=100\%$, 95%, and 90%). The polyadenine tracts were not required to be pure but were allowed to possess 0, 1, or 2 point mutations in the $h=100\%$, 95%, and 90% classes, respectively. This requirement makes flanks and repeats compatible as far as the mutational history is concerned. To obtain an impression about the form of the replacement frequency matrices, we present them graphically in Figs. 1–3. One can notice two distinctive features: (i) the $\tilde{A}^{(h)}(m,n)$ values decay roughly exponentially as a function of the absolute value of the difference between the two indices, and (ii) the repeats in lower mutual homology classes exhibit slower decay.

A. Replacement frequency and mutation probability matrices

Let us present and discuss the algebra on the basis of which the data that are presented in Figs. 1–3 were analyzed. The length modifications of polyadenine repeats can be treated by means of the formalism of rate equations [15]. Let us consider a population $N(n)$ of repeats that are in dynamic

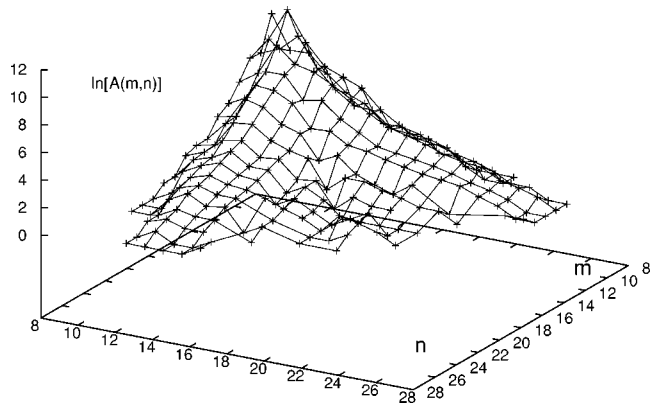

FIG. 2. Same as Fig. 1, except for the class $h=95\%$.

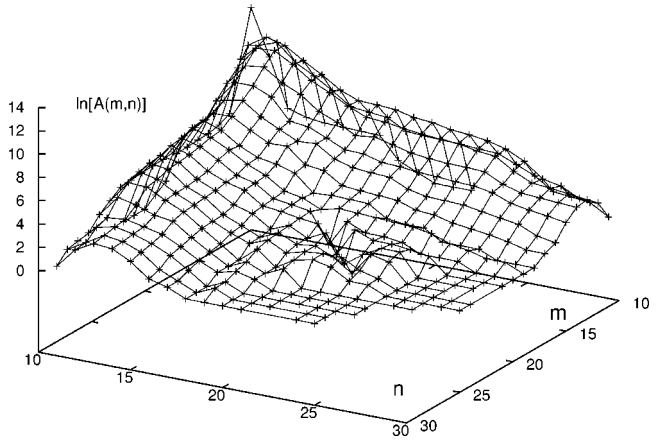

FIG. 3. Same as Fig. 1, except for the class $h=90\%$.

equilibrium. The repeat elements can be created, destroyed, or experience elongations and shortenings due to point mutations and slippage events. The rate equation for the repeat population $N(n)$ can be written as follows:

$$\frac{d}{dt}mN(m) = \sum_n w_{mn}nN(n) - mN(m)\sum_n w_{nm} \quad \text{for each } m. \tag{1}$$

The w_{mn} matrix element (belonging to the **w** matrix) represents the probability that a repeat with length n is transformed to a repeat with length m per unit of time. By means of the **w** matrix and $N(n)$ histogram one can make a quantitative prediction of the repeat modification events. The number of counts predicted by the theoretical model can be expressed as the product of the matrix element w_{mn} that represents the probability of the event and the "mass" $[nN(n)]$ of the repeats that can be subjected to the transition. However, since one has no control over the directionality of the transitions, it makes sense to take the average value of the $m\rightarrow n$ and $m\leftarrow n$ transitions, and we define the model replacement frequency matrix as follows:

$$A^{(t)}_{\mathrm{mod}}(m,n) = [W_{mn}(t)nN(n) + W_{nm}(t)mN(m)]/2. \tag{2}$$

The $\mathbf{W}(t)$ matrix is derived from the **w** matrix as described in the Appendix. The t parameter measures the time that elapsed from the moment when the pairs of repeats that are compared bifurcated and started their independent mutational history. According to the molecular clock hypothesis the homology of a pair of sequences that have a common origin decays as a function of time by the Jukes and Cantor formula [16] $h=h_0+(1-h_0)\exp(-p_m t)$, where h_0 is the homology of unrelated sequences and p_m is the probability for a point mutational event per nucleotide per unit of time. In the case of flanking regions h_0 is $1/4$ because of four different nucleotides participating in the sequence while within the mononucleotide repeats one can consider the sequence as being composed of two types of nucleotides: those belonging to the repeat (adenines in our case), and mutated ones, thus leading to $h_0=1/2$.

The matrix elements that are responsible for the creation and destruction processes are located in the first column and the first line of the **w** matrix, respectively. It is very difficult to parametrize the probabilities of these processes. One would need many parameters that are nearly impossible to determine and the only reasonable strategy is to devise a model where the creation and destruction processes are not taken explicitly into account.

A point mutational event that occurs within the repeat transforms a repeat with length n into two repeats with lengths n_1 and n_2 where $n_1+n_2+1=n$. The parametrization of mutational events is rather simple: One supposes that any site within the repeat is hit by the mutation with equal probability p_m.

The elongation and shortening processes are usually parametrized with one or two parameters. The one-parameter expression for the probability per unit of time that a repeat is subjected to the slippage event is usually written in the form $w_{mn}^{(s)}=np_s$. This implies that the slippage probability is proportional to the repeat length. Another possibility is to allow the slippage process to be unsymmetric [13] in the sense that elongations are not necessarily equally as probable as shortenings. In such a case one has to introduce another parameter that measures the asymmetry. In what follows we shall work out a methodology that will focus on symmetric slippage events. Ignoring the creation and destruction processes and under the hypothesis that the slippage process leads to repeat elongation with the same probability as for repeat shortening, the mutation probability matrix looks as follows:

$$w_{mn}=\begin{cases} nf(|m-n|), & n=m+1 \text{ or } n<m, \\ 4mp_m/(n-1)+nf(|m-n|), & n>m+1, \\ -\sum_{k(\neq m)} w_{km}, & n=m. \end{cases} \tag{3}$$

The function $f(|m-n|)$ defines the distribution of the probabilities of slippage step lengths, which broadens the standard supposition that is based on the model according to which the slippage step lengths are of unit size ($|m-n|=1$). The slippage step length distribution function $f(|m-n|)$ was assayed using three functional forms: a rectangular one, an exponentially decaying form, and an inverse power law form. The first term in the middle line on the right-hand side of Eq. (3) refers to the point mutational events that occur when a repeat of length n is split into two pieces with lengths $m=m_1$ and m_2. To focus attention on the slippage process we conduct the procedure by ignoring the point mutational events, that is, by neglecting the $4mp_m/(n-1)$ term. This means that we treat the repeats as if they were not interrupted by point mutations. In this way the concept of the molecular clock can be invoked, since the $\tilde{\mathbf{A}}^{(h)}$ data which are partitioned into three homology classes with $h=100\%$, 95%, and 90% are supposed to belong to three time intervals with gradually increasing amount of superimposed mutational events.

The $\mathbf{A}_{\mathrm{mod}}^{(t)}$ matrix can be compared with the replacement frequency matrix obtained by means of analyses of human genomic DNA sequences. In fact, the $\tilde{\mathbf{A}}^{(h)}$ matrices presented in Figs. 1–3 are not the most suitable quantities to be compared with $\mathbf{A}_{\mathrm{mod}}^{(t)}$, and we introduced the mass weighted $\mathbf{A}^{(h)}$ matrices whose matrix elements are proportional to the number of monomers that are involved in the slippage process. When comparing two repeats, one with length m and the other with length n one does not know whether the slippage direction is $m\rightarrow n$ or $m\leftarrow n$. The number of monomers that are subjected to the slippage process could be thus m or n, and therefore we defined a lower triangular mass weighted replacement frequency matrix as the average value:

$$A^{(h)}(m,n)=(m+n)\tilde{A}^{(h)}(m,n)/2. \tag{4}$$

To understand and reproduce the slippage process we should determine the parameters introduced into the mutation probability matrix $\mathbf{W}(t)$. The model replacement frequency matrices $\mathbf{A}_{\mathrm{mod}}^{(t)}$ that result from $\mathbf{W}(t)$ should fit to the natural replacement frequency matrices $\mathbf{A}^{(h)}$ that were generated by counting the pairs of repeats in homologous environments.

III. RESULTS

The numerical results show that the inverse power law leads to the best agreement between model and natural replacement frequency matrices. The slippage step length distribution function was modeled in the form $f(|m-n|)=p_s|m-n|^{-\alpha}/\zeta(\alpha)$. The parameter p_s is a measure of the probability that a slippage occurs. The factor $\zeta(\alpha)$ is the Riemann zeta function and takes care of the normalization of the step length distribution. When $\alpha\rightarrow\infty$ the distribution function approaches the δ function at $|m-n|=1$, which corresponds to the standard one-nucleotide slippage regime, while finite values of α mean finite width of the step length distribution. Since the disruption of the repeats by point mutations was only used as a measure of time that elapsed from the moment when the two sequences were duplicated, the p_m parameter does not enter into the construction of the model replacement frequency matrices. There remain only two parameters to be determined: the exponent α and the dimensionless (expressed in units of the inverse $1/p_s$ parameter) time parameter t for each homology class. A two-parameter fitting procedure does not represent a serious numerical task provided that the value function which is subjected to the optimization procedure is well defined and smooth enough. In our case the most straightforward choice of the quantity to be minimized would be an arbitrary norm of the difference between natural and model mass weighted replacement frequency matrices. We decided to use the absolute value of the difference between the logarithms of the first few columns of the $\mathbf{A}^{(h)}$ and $\mathbf{A}_{\mathrm{mod}}^{(t)}$ matrices, respectively. The logarithmic measure was used instead of a linear measure in order to attain the proper weight of far-from-diagonal elements that would have negligible weight in a linear scaling regime due to rapid decay of the $A^{(h)}(m,n)$ elements as a function of growing difference between the two indices. The fitting procedure was carried

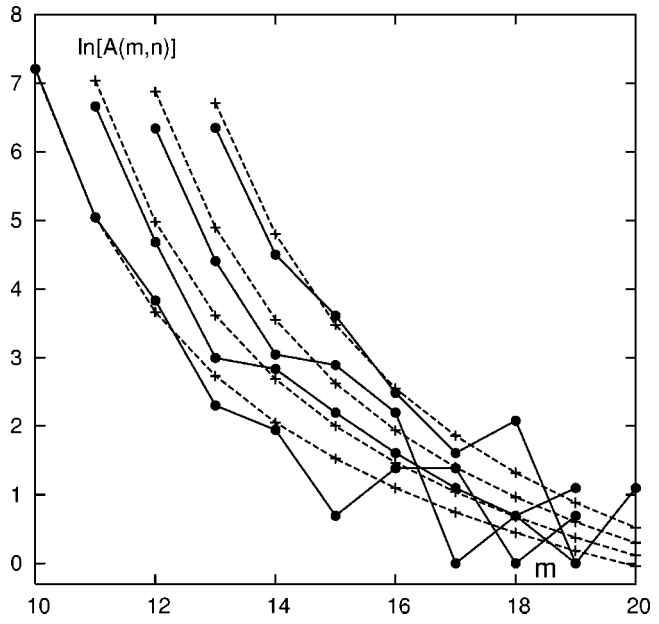

FIG. 4. The decay of columns of $\mathbf{A}^{(h)}$ (full lines) and $\mathbf{A}^{(t)}_{\mathrm{mod}}$ (dashed line) matrices. $\mathbf{A}^{(h)}$ matrix elements were deduced from natural DNA sequences and they are compared with $\mathbf{A}^{(t)}_{\mathrm{mod}}$ which is the result of our model. $h=100\%$.

out by a simple brute force search for the pairs of α and t values that produce optimal fit between $\mathbf{A}^{(t)}_{\mathrm{mod}}$ and $\mathbf{A}^{(h)}$. The following results were obtained: $\alpha=2.4\pm0.4$; $p_s t=0.03$, 0.05, and 0.07 for the 100%, 95%, and 90% homologies in the flanking regions, respectively. In Figs. 4–6 the counts performed on natural sequences and the results of our model are compared. Full lines represent the columns of mass weighted replacement frequency matrices drawn on a logarithmic scale. Dashed lines represent the results of our model with the parameters fitted to reach the optimal agreement with the natural matrix counterpart. The fitting procedure was hindered due to strong scatter of $\mathbf{A}^{(h)}$ elements, but in spite of that we think that the agreement between the matrices deduced from natural DNA sequences is satisfactory.

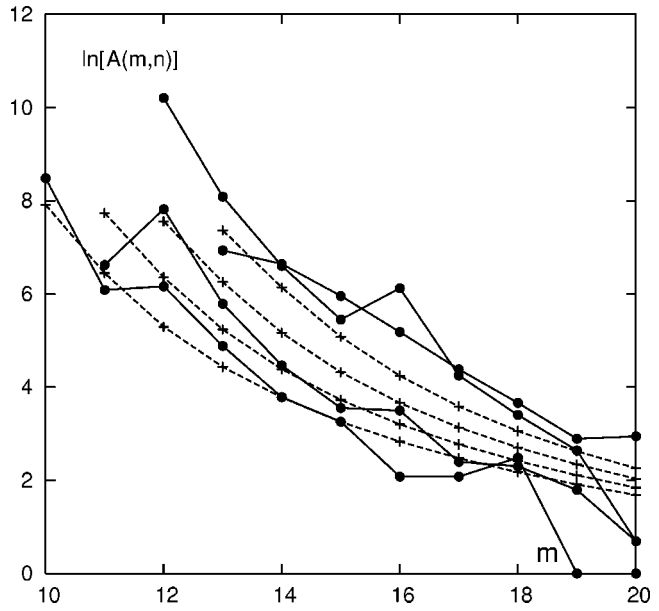

FIG. 5. Same as Fig. 4 for class $h=95\%$.

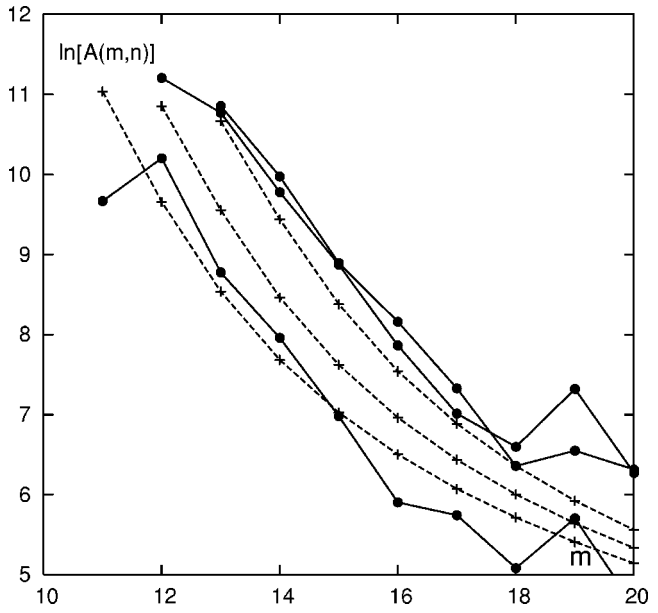

FIG. 6. Same as Fig. 4 for class $h=90\%$.

IV. DISCUSSION AND CONCLUSIONS

In this work we demonstrate that the mutational dynamics of short sequence repeats can also be treated in a standard way by searching for pairs of highly homologous sequences and inferring the repeat modification events from the observed differences between the pairs of sequences that are supposed to have a common ancestor. In our case we deal with three types of mutational events: repeat elongations and shortenings, point mutations in the regions of repeats, and point mutations in the flanking regions. Algebra based on the master equation formalism enabled us to resolve the problem of superpositions of mutational events and to single out the probabilities of elementary mutational events.

Our results enable us also to draw a comparison between point mutational and slippage processes. The three homologies (100%, 95%, and 90%) correspond, according to the Jukes-Cantor formula, to the values $p_m t=0$, 0.069, and 0.14. Let us compare these values with the corresponding values pertaining to slippage processes. The first pair of values $p_s t=0.03$, $p_m t=0$ is consistent with the notion that slippage processes are running at a higher pace than point mutations. This notion is based on a scenario that proposes that slippage processes and point mutations contribute comparable amounts of change at each turnover of human generations. The values of the parameters that support this regime are $p_m=2.2\times10^{-9}$/yr/nucleotide [17] and one to two orders of magnitude higher value of p_s [14]. However, the second two pairs of $p_s t$, $p_m t$ values ($p_s^{95}/p_m^{95}=0.05/0.069$ and $p_s^{90}/p_m^{90}=0.07/0.14$) indicate that the balance between p_s and p_m need not be so much in favor of p_s. One should of course be aware that the comparison that is based on the alignment of

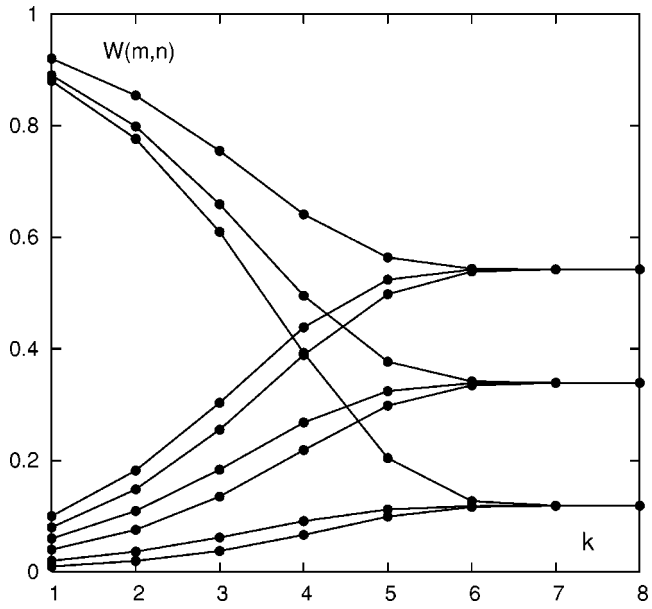

FIG. 7. Limiting behavior of $W(m,n)$ matrix elements predicted by Eq. (A4). The initial values of nine matrix elements of the 3×3 **W** matrix [Eq. (A5) with $a=0.1$, $b=0.08$, $c=0.06$, $d=0.04$, $e=0.02$, and $g=0.01$] are displayed along the vertical axis at the left side of the figure. The values of matrix elements at higher powers of the **W** matrix converge toward $f(1)=0.542$, $f(2)=0.339$, and $f(3)=0.119$ values that are displayed along the vertical axis at the right side of the figure. The variable k along the horizontal axis represents the binary logarithm of the power to which the matrix was raised. The uppermost three curves belong to the diagonal elements, while the remaining ones belong to the off-diagonal matrix elements.

flanks (with 95% and 90% homologies) of the repeats that contain one or two point mutations are treated as if they were pure repeats. This trick helped us to build the concept of the molecular clock directly into the slippage process, but on the other hand we introduced some inconsistency because mutated repeats do not necessarily exhibit the same slippage propensity as nonmutated ones and since the time ordering of the two events is unknown, our results corresponding to $h=95\%$ and 90% are to some extent uncertain. We should thus mostly rely on the $h=100\%$ case which does not provide us a quantitative measure about the p_s/p_m ratio, but only indicates that the slippage process runs faster than the point mutational process.

To conclude, let us recapitulate our findings. The main interest of our results is the evidence that it is possible to go beyond a simple model which assumes that the step length distribution in the slippage process has a trivial δ function appearance with $\Delta n=\pm1$. The results predict a monotonically decaying step length distribution in the form of an inverse power law, not very far from a δ function, since the distribution decays faster than the inverse second power. Only one-third of slippages are predicted to go beyond the single nucleotide step length.

ACKNOWLEDGMENTS

The financial support of the Ministry of Education, Science and Sport of the Republic of Slovenia is gratefully acknowledged. Norma Mankoč Borštnik is acknowledged for helpful discussions regarding the properties of w matrices.

APPENDIX: THE PROPERTIES OF THE w MATRIX

In this appendix the properties of the **w** matrices are discussed. For the stationarity condition $d[N(m)]/dt=0$ Eq. (1) acquires the form $[\mathbf{w}]\mathbf{f}=0$ with $\mathbf{f}=n\mathbf{N}(n)/\Sigma nN(n)$ and $w_{mm}=-\Sigma_{k(\neq m)}w_{km}$. It is convenient to incorporate the dt parameter into the **w** matrix and to define

$$W_{mn}(dt)=\begin{cases} w_{mn}dt, & m\neq n,\\ 1-w_{mm}dt, & m=n. \end{cases} \tag{A1}$$

In this notation we obtain

$$[\mathbf{W}]\mathbf{f}=\mathbf{f}. \tag{A2}$$

This equation has the form of an eigenvalue problem with one eigenvalue being $\lambda=1$. In principle also the remaining eigenvalues and eigenvectors can be determined, but it is necessary to keep in mind that in spite of the fact that the **W** matrix is real, it is not necessarily symmetric and is thus non-Hermitian. The remaining eigenvalues are in general complex; also the eigenvectors are complex and not necessarily orthogonal. Numerical diagonalization of real nonsymmetric matrix is a tedious procedure, and it is hard to obtain proper numerical routines that solve the problem. However, the $\lambda=1$ eigenvector of our **W** matrix is an exception because, since the eigenvalue is known, one can apply the subroutine for the solution of the system of linear equations with the Gauss-Jordan elimination procedure and determine the $\lambda=1$ eigenvector.

By means of the **W** matrix one can express the time variation of the **f** vector in the following way:

$$\mathbf{f}(dt)=[\mathbf{W}(dt)]\mathbf{f}(0) \tag{A3}$$

This equation is valid for short enough time intervals which means $p_s dt\ll1$ and $p_m dt\ll1$. To propagate the composition vector for a longer time period, say $t=k\,dt$, one has to operate on $\mathbf{f}(0)$ with $\mathbf{W}(dt)$ raised to the kth power: $[\mathbf{W}(t)]=[\mathbf{W}(t/k)]^k$.

The matrix $[\mathbf{W}(t)]$ possesses the following limiting property:

$$\lim_{k\to\infty}[\mathbf{W}(dt)]^k=[\mathbf{f},\mathbf{f},\ldots,\mathbf{f}] \tag{A4}$$

where **f** is the eigenvector of $\lambda=1$ eigenvalue. Let us illustrate this for the three-dimensional case. A general form of the matrix looks as follows:

$$\mathbf{W}=\begin{vmatrix} 1-c-e & a & b\\ c & 1-a-g & d\\ e & g & 1-b-d \end{vmatrix}. \tag{A5}$$

For our purpose all six parameters are supposed to be positive and to satisfy the condition $\Sigma=a+b+c+d+e+g<1$. The diagonalization of the matrix can be done by hand and one obtains $\lambda_1=1$ and $\lambda_{2,3}=1-\Sigma/2\pm\sqrt{\Sigma^2/4-4S}$ where $S=ab+ad+ae+cb+cd+cg+bg+de+eg$. Also the eigenvec-

tors can be calculated. We only give the one that corresponds to $\lambda_1=1$: $f(1)=(ad+ab+bg)/S$, $f(2)=(cb+cd+eb)/S$, $f(3)=(ae+cg+eg)/S$. In Fig. 7 it is shown how the matrix elements transform in the process of arising the matrix to higher powers. Also our **W** matrices that were used to define the model replacement frequency matrices for the three homology classes used in Sec. II are behaving in a similar way as the matrices whose matrix elements are exhibited in Fig. 7.

[1] E. S. Lander *et al.*, Nature (London) **409**, 860 (2001).

[2] B. Borštnik, D. Pumpernik, D. Lukman, D. Ugarković, and M. Plohl, Nucleic Acids Res. **22**, 3412 (1994).

[3] D. Metzgar, J. Bytof, and C. Wills, Genome Res. **10**, 72 (2000).

[4] G. Toth, Z. Gaspari, and J. Jurka, Genome Res. **10**, 967 (2000).

[5] B. Borštnik and D. Pumpernik, Genome Res. **12**, 902 (2002).

[6] R. Cox and S. M. Mirkin, Proc. Natl. Acad. Sci. U.S.A. **94**, 5237 (1997).

[7] S. H. Eom, J. Wang, and T. A. Steitz, Nature (London) **382**, 278 (1996).

[8] L.-C. Hsieh, L. Luo, F. Ji, and H. C. Lee, Phys. Rev. Lett. **90**, 018101 (2003).

[9] S. Kruglyak, R. T. Durett, M. D. Schug, and C. F. Aquadro, Proc. Natl. Acad. Sci. U.S.A. **95**, 10774 (1998).

[10] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).

[11] B. Borštnik, D. Pumpernik, and D. Lukman, Europhys. Lett. **23**, 389 (1993).

[12] D. Holste, I. Grosse, and H. Herzel, Phys. Rev. E **64**, 041917 (2001).

[13] B. Borštnik, and D. Pumpernik, Europhys. Lett. **65**, 290 (2004).

[14] G. I. Bell, Comput. Chem. (Oxford) **20**, 41 (1996).

[15] X. Gu and W.-H. Li, Proc. Natl. Acad. Sci. U.S.A. **95**, 5899 (1998).

[16] T. H. Jukes and C. R. Cantor, in *Mammalian Protein Metabolism*, edited by N. H. Munro (Academic Press, New York, 1969).

[17] R. H. Waterson *et al.*, Nature (London) **420**, 520 (2002).