Letter

Tandem Repeats in Protein Coding Regions of Primate Genes

Branko Borštnik[1,2] and Danilo Pumpernik[1]

[1]National Institute of Chemistry, SI-1001 Ljubljana, Slovenia

Tandem repeats in GenBank primate nucleotide sequences annotated as protein coding regions are analyzed. It is found that only trinucleotide repeats show repeat enrichment well above the threshold of statistical significance. The statistics are improved by a simultaneous search for repeats on both the amino acid and nucleotide levels. The results of the analyses of natural sequences are interpreted by comparing them with the results of the computer simulation of the model dedicated to protein coding regions. According to the simulation results, a limited set of trinucleotides, that is, cgg, ccg, cag, and gaa repeats coding for polyalanine, polyglycine, polyproline, polyglutamine, and polylysine are prone to proliferation. It is also found that within the repeat regions slippage is more frequent by a factor of 10 than point mutations, whereas the ratio of silent versus recognizable point mutations is approximately the same as elsewhere in coding regions. The trinucleotide repeats cover slightly more than 0.3% of the protein coding regions of genes.

Tandem repeats with short (1–6 bp) monomer units, also called microsatellites or simple sequence repeats, exist in noncoding genomic regions as well as in regions coding for proteins and structural nucleic acids. In the human genome (Genome Sequencing Consortium 2001) approximately 2% of the nucleotide sequences are in the form of tandem repeats in which the length of the repeat unit is between 1 and 11 bp. The functional role of tandem repeats is poorly understood. They are, however, known to be involved in several genetic diseases and they can be successfully used as the genetic markers. To shed more light on the character of these short sequence repeats, it is worthwhile to determine the type and content of short sequence repeats in coding regions and to make a quantitative comparison with the situation in noncoding regions. Three research groups are involved in just such an effort (Toth et al. 2000; Metzgar et al. 2000; Field and Wills 1998). They have shown that short tandem repeats are much more numerous in noncoding regions than in protein coding regions, and that in coding regions, trinucleotide and hexanucleotide repeats are more frequent than mononucleotide, dinucleotide, tetranucleotide, and pentanucleotide repeats. This latter finding can be explained in terms of the complete impairing of the protein function by the frameshift mutation that takes place when an exon receives insertion or deletion of a segment whose length is not a multiple of codon length.

The major problem with analyzing coding regions is the scarcity of annotated sequences. To better determine the content and the characteristics of microsatellites in exons we have made two kinds of improvements in the analyses: (1) The counts of tandem repeats in exons were performed on the largest data set possible; and (2) analyses of the counts were interpreted according to a realistic expectation model supported by a computer simulation.

Concerning the first point, it is clear (especially now, because the results of the initial sequencing of the human genome are available) that the primate sequences represent the largest reservoir of sequences concentrated in a small fraction of the phylogeny. However, because our interest is to quantify the differences between coding and noncoding regions, only the entries with annotated locations of exons are acceptable. Nucleotide databases contain approximately 10^3 protein products whose functions are well known and for which there may even be a known three-dimensional structure. At the other extreme, some annotations have been performed with computer programs that have a very low degree of fidelity. In between there are a number of experimental techniques and procedures by which the certainty that a particular DNA segment belongs to an active coding region is determined with varying degrees of fidelity.

The second point refers to models that can be used to explain the appearance of tandemly repeated sequences. The simplest is the Bernoulli model, which is based on the expectation that any sequence is present in the genome with a probability based on the average DNA composition (Cox and Mirkin 1997; Metzgar et al. 2000). More complex models are based on the Markov chain, random walk, or birth and death (birth and death [BD]) approaches (Kruglyak et al. 1998; Bell 1996). We shall elaborate on a model that, because of its complexity, does not possess an analytical solution. We have designed a computer simulation procedure based on a model that is a variant of the standard models adapted for protein coding regions. By comparing the results of the computer simulation with the results of the analyses of natural sequences, one can get useful information about the nature of repeat generation in protein coding regions.

The protein coding regions represent an essentially different environment to the repeats than the noncoding regions. The main difference stems from the greater evolutionary constraints imposed on protein coding regions by protein function compared with the constraints of an as yet unknown origin that are imposed on noncoding regions. Accordingly, repeats in coding regions are strongly suppressed and it is hard to obtain statistically significant data. Furthermore, the role of point mutations is more complex in coding regions than elsewhere because of the fact that there are two classes of point mutations: silent (interchange between two synony-

[2]Corresponding author.
E-MAIL branko@hp10.ki.si; FAX (386-1) 4760-300.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.138802.

mous codons) and recognizable (leading to amino acid replacement), which have a different frequency. Therefore, a reliable model must be flexible yet complex enough to cope with all of the above mentioned features.

In our earlier work (Borštnik et al. 1993) we showed that there is a substantial difference between coding and noncoding DNA regions in terms of the nucleotide–nucleotide correlations. We have shown an exponential decay in the correlation functions observed in exons, whereas long range correlations appear to exist in noncoding regions. A key difference between the repeats in coding and noncoding regions is that in coding regions a trinucleotide repeat (such repeats are the central point of this study) could have any of three possible interpretations, depending on the reading frame that is imposed on it.

In this paper we limit ourselves to mononucleotide, dinucleotide, and trinucleotide repeats. The essential advantage that we want to exploit is the ability to detect, register, and analyze repeats on both the nucleotide and amino acid level. It turns out that it is exactly this detail, namely, the accessibility to the longer runs present in amino acid sequences, that helps us to make quantitative conclusions about the nature of the distributions of the repeats.

RESULTS

Mononucleotide and Dinucleotide Repeats

Our results on the frequency of appearance of tandem repeats of mononucleotides and dinucleotides in coding regions are presented in Figures 1 and 2. In both cases the histograms decay at repeat lengths less than 12 nucleotides (nt) with the rates predicted by the Bernoulli model. The runs of mononucleotide and dinucleotide repeats whose length exceeds 12 nt show slower decay, which has been termed "repeat enrichment" in the literature (Cox and Mirkin 1997). However, as we can see in Figures 1 and 2, the repeat enrichment is rather poorly expressed. Mononucleotide repeats coding for polyglycine ($g_{3n} = Gly_n$), polyproline ($c_{3n} = Pro_n$), polylysine ($a_{3n} = Lys_n$), and polyphenylalanine ($t_{3n} = Phe_n$) have been found, but all the above mentioned homopolypeptides can be also coded by other heteronucleotide codons. It turns out that there is a tendency to avoid homonucleotide tracts in coding regions: The statistics show that homonucleotide tracts are underrepresented in polyglycine, polyproline, polylysine, and polyphenylalanine coding. A similar situation has been encountered in the case of the coding of tandemly repeated dipeptides that are most frequently coded by nonperiodic nucleotide sequences rather than by tandemly repeated dinucleotides.

Trinucleotide Repeats

Trinucleotide repeats in exons produce runs of identical amino acids. Let us take alanine repeats as a specific example. One can find four types of repeats coding for alanine : $(gca)_n$, $(gcc)_n$, $(gcg)_n$, and $(gct)_n$. The alanine stretches are subjected to two kinds of point mutations: recognizable ones, which occur when any codon is changed at the first or second position, and silent mutations, when the third position in a codon is changed. In addition, according to the slippage hypothesis, stretches may also be subjected to elongation and shortening processes. Analysis of the exon sequences can reveal, to some extent, their mutational history. Histograms can be used to show the frequency of appearance of repeats of each codon, as well as the distribution of alanine stretches. The latter type of histogram can show the existence of longer runs on amino acid level, because silent mutations, which fragment runs of identical codons on the DNA level, do not affect the final amino acid content.

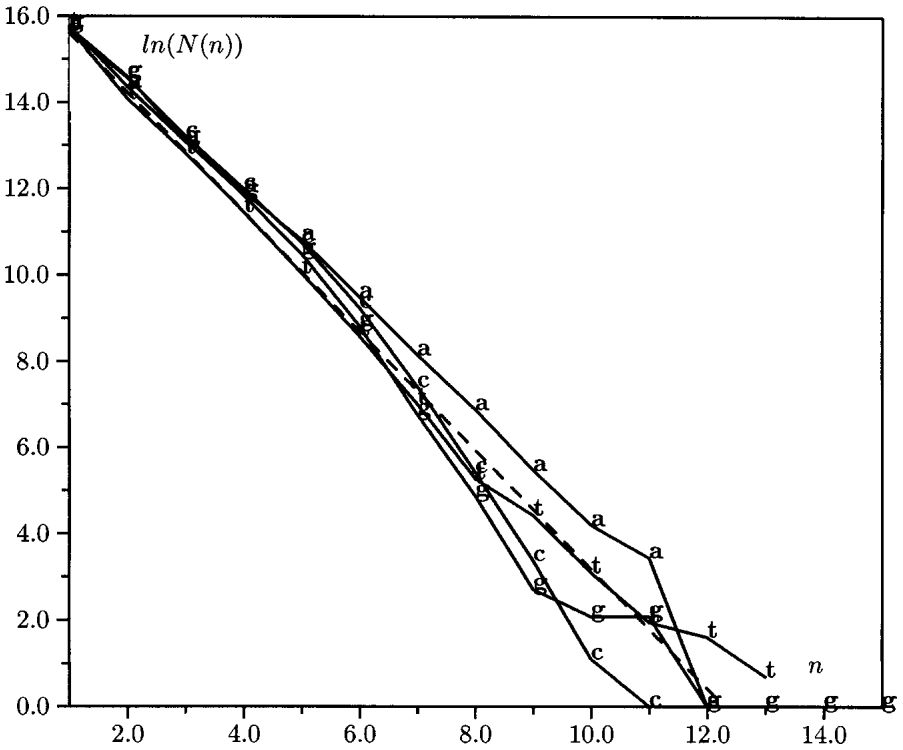

Figure 1 Four histograms displaying the frequency of a, c, g, and t mononucleotide repeats in exons. The dashed line indicates the slope of the histogram of random sequences with equiprobable nucleotide compositions.

In Figure 3, the five alanine histograms, which we generated by analyzing the coding sequences of primates (GenBank, release 122), are depicted as full lines. The uppermost histogram refers to the distribution of runs in amino acid representation. The four histograms corresponding to four alanine codons are labeled with a, c, g, and t letters. The following two features are important: The well-resolved crossover of two regimes in the alanine histogram at $n = 5$ and the short range of all four histograms corresponding to the four alanine individual codons that only marginally exceed the limit of the exponential decay of codons according to their frequency of appearance. It is thus evident that we have to build up our strategy for understanding tan-

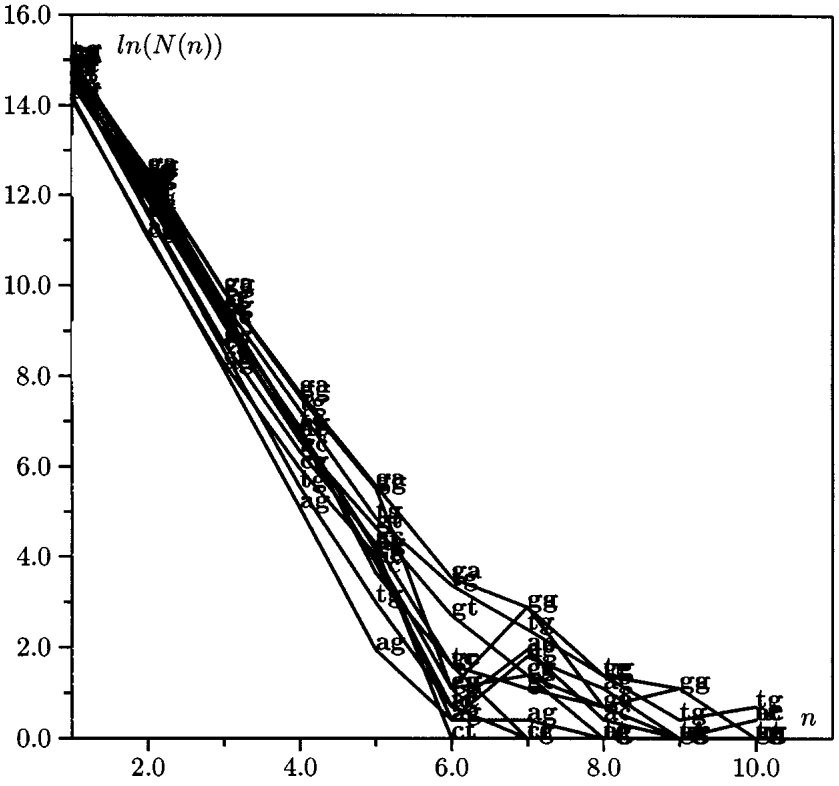

Figure 2 Histograms displaying the frequency of 12 heterodinucleotide repeats in exons. At the *bottom right* part of the diagram, a minimal degree of repeat enrichment is observed. The *upper left* part of the diagram shows the standard decay of the frequency of dinucleotides as predicted by the Bernoulli model.

dem repeats in exons with data provided by the histograms based on the counts of amino acid repeats. Data on the enrichment of individual codons should be used to complement the data on amino acid enrichment.

The uppermost two histograms in Figure 3 (the results of the analyses of natural sequences [full line] and the results of the computer-simulation procedure [dashed line]) that show the frequency of the appearance of alanines in semilogarithmic plot (*log N*(*n*) as a function of *n,* where *N(n)* is the number of alanine runs with length *n*) can be fitted with two straight lines (dotted lines in Fig. 3), one below and one above the $n = 5$ value. There are two alternative possibilities for the interpretation and quantification of the repeat enrichment phenomenon.

Effective Codon Probabilities

The negative slope of the alanine histogram in Figure 3 at $n < 5$ corresponds to the average content of alanine, whereas the other four histograms reflect the usage of the four codons. The values conform to the available data on codon usage (Nakamura et al. 2000). From the $n > 5$ part of the uppermost histogram, where the slope has a lower value, one can extract the effective alanine composition of the repeat regions, p_r(Ala) = 68%. This means that within the alanine repeat regions, one-third are nonalanine codons. The location of the crossover point and the slopes of straight lines that try to fit the uppermost histogram of Figure 3 enable us to determine the extent of alanine repeats in the protein coding regions. Quantitative measures of the alanine repeat coverage can be obtained from equation M1 and results in 0.05% ($=L_r/L_O$) where L_r is the length of alanine repeats and L_O is the total coding sequence length. The values of the coverage of the repeats of other amino acids is given in Table 1. The values are in agreement with the findings that we obtained by direct counts of the repeats. The values are also close to the results in the papers of Toth (2000) and Metzgar (2000); their values are in the range between 0.1% and 0.2% of the trinucleotide repeats in exons.

Interpretation via a Computer Simulation Procedure

Our simulation model, which is described in the Methods section, is compatible with the idea of two densities, as shown by the two values of the slope in Figure 3. Along the major part of the sequence where the repeats are not enriched, the point mutations generate short repeats. Repeat insertions, elongations, and shortenings, which are introduced within the framework of the simulation procedure, are responsible for the second part of the diagram with a low value for the negative slope.

The calculation of the long range behavior of the four histograms for the four alanine codons was impaired to some extent by insufficient statistics, but we were still able to fit them using the results of our simulation procedure. We determined that the mechanism that generates alanine repeats produces predominantly (gcg) repeats, with approximately 11% probability for (gct) repeats and 3% probabilities for (gca) and (gcc) repeats. The relative uncertainty of the above mentioned probabilities is high because of the strong degree of scatter in the enriched regions of the histograms belonging to the individual codons. The error was estimated in the course of the simulation procedure. The α_c parameter (see Methods section) was determined with approximately 30% uncertainty concerning what produces the errors that are depicted in Table 1.

In addition to alanine, the analysis was also performed for all the amino acids whose coding triplets show an above average tendency for repeat formation. In Table 1 we present the parameters that resulted from the simulation of alanine, glycine, proline, glutamine, and lysine repeats. It turns out that the (gcg) codon that is responsible for the most frequent alanine repeats also produces the most frequent glycine repeats when being read in another reading frame as $(ggc)_n$. The same repeat being read in the third possible reading frame as $(cgg)_n$ contributes to the arginine repeats, which we did not quantify. The next element that is frequently encountered is the complement of (ggc) and codes for proline repeats $(Pro)_n = (ccg)_n$. Two other proline codons, (cca) and (cct), contribute approximately three times less frequently to proline repeats, whereas our parameterization (Table 1) does not predict pure $(ccc)_n$ as a source of proline repeats.

It is important to note that there is no correlation between codon usage and codon propensities for the repeat formation. In the case of the alanine (gcg) codon, the two quan-

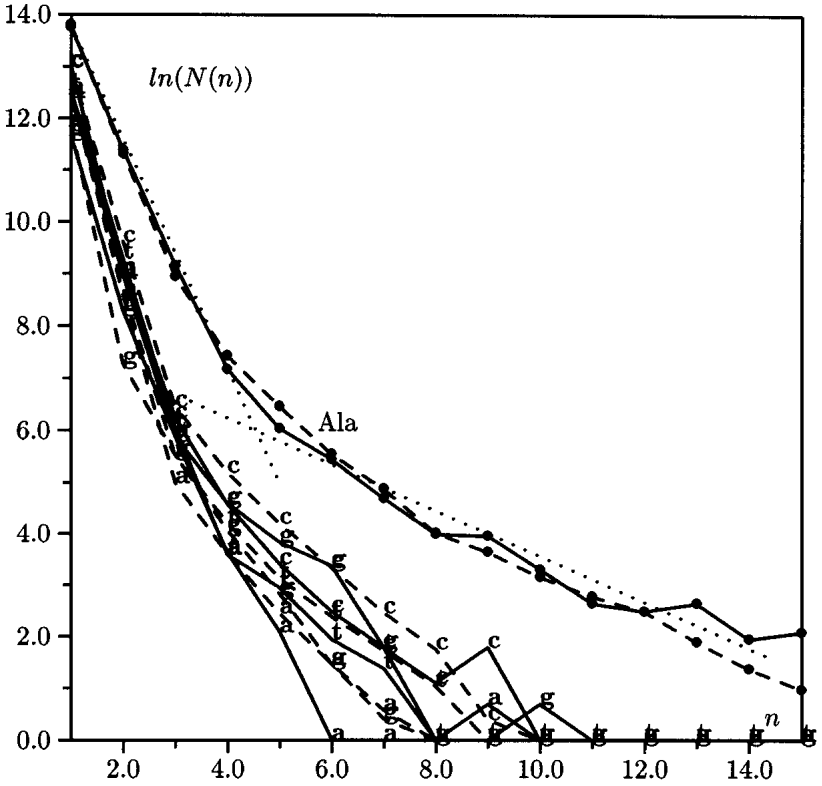

Figure 3 Histograms representing the abundance of runs of the four alanine codons and of alanine itself. Full lines represent the results of the counts performed on natural sequences; dashed lines are the simulation results. The *uppermost* histogram belongs to the alanine runs on the amino acid scale. The two dotted lines represent the linear fit of the two parts of the alanine histogram. The letters a, c, g, and t mark the histograms of gca, gcc, gcg, and gct codons, respectively. Note that the gcg codon, which only accounts for 10% of alanines, is accordingly low and descends steeply at n < 5, only becoming the most persistent at high n values.

tities are negatively correlated: this codon has the lowest codon usage (its share among the four possible codons is 1/9 instead of 1/4) but the highest propensity for repeat formation (Table 1).

In regard to glutamine and lysine repeats, it is important to note that in both cases only one of the two codons is responsible for the existence of runs. In the case of glutamine repeats, the dominance of (cag) over (caa) is much more clearly expressed than in the case of lysine, where $(aag)_n$ insertions are only approximately three times more frequent than $(aaa)_n$ insertions.

The results clearly indicate that in coding regions homonucleotide tracts are suppressed and that homopolypeptide tracts that could be coded by homopolynucleotide tracts show a codon usage that avoids aaa, ccc, and ggg codons. This assertion is supported by Figures 1 and 2, which show that we were not able to detect the repeat enrichment in coding regions.

Mutational Dynamics

There is no way of determining absolute mutation rates, but we can get an idea of the relative rates. Three mutation rates can be compared: the rate of repeat insertions, the rate of elongation/shortening (the slippage rate), and the rate of point mutations at recognizable and silent sites. We decided to keep the number of parameters as limited as possible. To this end we used the same rate parameter for repeat elongation and for the shortening events. The ratio between silent and recognizable mutations was determined to have a value between 5 and 10, which means that silent mutations are 5 to 10 times more frequent than recognizable ones. These values are in the range usually quoted in the literature.

The most important results are the ones that refer to the ratios of various mutational events, such as point mutations, repeat insertions, and slippage events. According to the literature data (Kruglyak et al. 1998; Bell 1996), the rate of neutral point mutations is 10^{-7} per base pair per generation for primates and the slippage probability is 10^{-3} per repeat per generation, which means that if one takes an average repeat length somewhere between 10 and 100 bp, the ratio between slippage events versus point mutations is between 100 and 1000. In the case of coding regions, the results may differ because greater constraints imposed by protein functions reduce the rate of both point mutations and slippage events, although it is not known in what relative proportion. Our results indicate that slippage events are only approximately tenfold more frequent than point mutational events.

DISCUSSION

The fact that in human genes the most frequently encoun-

Table 1. The Relative Probabilities of Elongation/Shortening Events for the Runs of Codons That Code for the Respective Amino Acids

	Ala(gcx)	Gly(ggx)	Pro(ccx)	Gln(cax)	Lys(aax)
p_a	0.03 ± 0.02	0.10 ± 0.05	0.20 ± 0.10	0.10 ± 0.05	0.30 ± 0.10
p_g	0.83 ± 0.10	0.05 ± 0.03	0.60 ± 0.10	0.90 ± 0.10	0.70 ± 0.10
p_c	0.03 ± 0.02	0.75 ± 0.10	0.05 ± 0.03		
p_t	0.11 ± 0.05	0.10 ± 0.05	0.20 ± 0.10		
L_1/L_0	0.0005	0.0008	0.0005	0.0008	0.0002

The codons are identified by their third base (p_a of alanine refers to [gca] codon). In the bottom row the relative coverage of exons by the repeats are given. The estimated relative error of L_1/L_0 values is 50%.

tered polymorphism is the repeat length polymorphism and that it has its roots in repeat elongation/shortening events indicates that such processes are important ingredients of molecular evolution. In the final consequence, the repeat elongation/shortening processes also lead to the increase of biological complexity, which is considered to be a hallmark of biological evolution. Repeats in nucleotide sequences have been linked to the notion of molecular parasitism (Borštnik et al. 1994), but there is also evidence (Nishizawa 1999) that a major part of the protein coding region has evolved through the combined mechanisms of duplication and substitution. This paper attempts to present a methodology that will provide a better quantitative assessment of the statistics of short repeats in the coding regions of genes. The percentage of trinucleotide repeats in exons, for example, diverge by nearly a factor of two if we compare the results of different investigators. We have shown that it is possible to improve the statistics by combining the repeat counts at both the DNA and amino acid level. We can also improve the statistics by discriminating between the repeats that occur by chance (contributing to $N(n)$ at small n value) and the high n value repeats that appear as a result of standard DNA amplification mechanisms such as slippage replication or unequal crossing-over (Dover 2000).

Our methodology draws attention to the fact that several questions still remain unanswered, such as exactly which trinucleotides are participating in the slippage events. It appears that trinucleotide proliferation has its roots on the DNA level, but it is not clear whether is it a pure coincidence that the most frequent repeats are coding for the frequently occurring amino acids. The list of the most probable candidates for proliferation emerging from our analyses is compatible with the results of other investigators. However, it is still unclear whether such proliferation occurs over a limited set of trinucleotides or if all trinucleotides can proliferate to varying degrees. Silent point mutations are continually converting the repeats to other forms. Only analyses based on sufficient statistics and proper methodology can provide the most reliable answers to the question of the role of tandem repeats in the genome and in its evolution.

METHODS

Analyses of Primate Sequences

The searches for tandem repeats were performed on our own computer. GenBank (Benson et al. 2000) nucleotide sequences of primates (release 122) were transferred to our computer and were analyzed with computer programs prepared for the purpose. The outline of the computer program is the following: The GenBank entries were checked for the presence of coding regions, which were then searched for the presence of tandem repeats. The length of the repeat monomers was allowed to be up to 12 nt, but because trinucleotide repeats are the most informative, we focused our attention on mononucleotide, dinucleotide, and trinucleotide repeats. In the case of trinucleotide repeats, the analyses were performed to calculate the histograms corresponding to the runs of each individual codon and also to calculate the distribution of the runs of the amino acid itself.

Substantial effort was devoted to the process of selecting GenBank entries to avoid the use of too closely homologous sequences and sequences whose coding regions are annotated with insufficient fidelity. Among the GenBank entries are many sequences whose exons are described as *putative, unnamed* or *unknown protein product, hypothetical, similar to*. We put such sequences aside and analyzed them separately, but we found that the results obtained by analyzing such sequences did not differ substantially from the results obtained by processing the group of sequences without the previously mentioned remarks.

The Expectation Model and Computer Simulation Procedure

The most primitive model is the Bernoulli model. According to this model, the probability of existence of a particular sequence is expressed as a product of the probabilities of the presence of the nucleotides of which the sequence is composed. The Bernoulli model is based on the premise that DNA sequences are the results of the fixation of randomly generated point mutations of a predominantly neutral character. The mechanisms that generate tandem repeats are far from this scenario. The most probable cause of tandem repeats lies in the slippage mechanism, either at the recombination or DNA replication stage. The expectation model should inevitably conform to such a scenario. The literature contains several models that have been constructed to explain the abundance of microsatellites in various genomes (Kruglyak et al. 1998; Bell 1996). They are usually based on methodologies that are used to describe stochastic processes (Markov chain, random walk, birth and death processes). The models can be divided into two classes: They may be or may not be amenable to exact solution. The essential ingredients of the models are the measures of how and under what conditions the repeats are elongated/shortened, how new repeats are introduced, and how point mutations interfere with the repeats. In the most compact form, the model can contain only one adjustable parameter (Kruglyak et al. 1998): the ratio of slippage versus the point mutation rate. The above mentioned models do not cope with the specific conditions that are imposed in coding regions, in which the repeats can have two or three different meanings in different reading frames and two different kinds of point mutations can take place: silent and recognizable. To take into account these features, we had to design a new model that cannot be maintained at such a low level of complexity that it would still be analytically tractable. The existing models of other investigators were therefore not suitable for our purpose.

Our model is an extension of the models of the authors cited above. The basic elements of the model are the following alterations of nucleotide sequences: repeat insertion/elongation/shortening and two types of nucleotide point mutations (silent and recognizable point mutations). There are several ways to examine the repeat insertions. The most straightforward option is to choose the repeat "seeds" from among the runs of identical trinucleotides that appear accidentally. One needs to decide on the minimal acceptable length of the repeat and the probability that a particular repeat would be chosen as the seed for an elongation/shortening procedure. For each amino acid, one should have a set of parameters, $p_{n,c}$ where n refers to the repeat length and c to the codon identity. It is obvious that $n = 2$ is the minimal repeat length that can play the role of the seed of the repeat. The interval of seed lengths does not extend far beyond this minimal value because $n = 3$ cases are already rare, whereas $n = 4$ cases (four subsequent identical codons) are expected only with the probability 1/16 per Mbp and their rate of formation is too slow to compensate for the decay of the repeats as a result of shortening events. To keep the model as simple as possible, we worked with $n = 2$ seeds only.

The implementation of elongation/shortening and point mutational events needs several inputs: the probabilities that a particular repeat will be modified, the probability for elongation versus shortening, and the number of monomers to be added or removed, as well as the rate of recognizable and silent mutations.

We prepared a computer program to simulate the evolu-

tion of tandem repeats within exons. The simulation was performed on the model exon (which can be also treated as a series of several exons) with a length of 500,000 codons. Each of the five amino acids (Ala, Gly, Pro, Gln, and Lys) that were the subject of our study were treated in separate runs. The nucleotide sequence was initialized as a random sequence with all 4 nt equally probable. In the next step, the sequence was upgraded to attain the state in which the sequence would comply with the primate codon usage of the amino acid whose repeats are processed. After this phase is accomplished and the sequence displays proper codon usage, the process of repeat formation and repeat elongation starts. In consecutive cycles the following operations are repeated: The pairs of tandemly repeated codons of a certain type are identified as the repeat seeds (with the probability p_{2c}, where c stands for the codon identification, for example, gca, gcc, gcg, and gct in the case of alanine). The length dependence of the probability for the elongation/shortening events is taken to be of the standard form (Kruglyak et al. 1998; BD of Bell 1996): $p_{e/s}(n) = (n-2)b$, where n is the number of repeating codons and b is the basic parameter referring to the rate of slippage. The elongation and shortening events are taken to be equally probable. In each cycle, point mutations were also performed. Each codon is subjected to a mutation, either a recognizable (probability p_{rec}) or silent point mutation. The silent mutations ($p_{s,c}$) of the codons that belong to the amino acid whose repeats are the subject of the simulation are performed in such a way that the codon usage is conserved. This is achieved by using mutation probabilities that are proportional to the usage of a codon resulting in a mutational process ($p_{s,c} \propto$ codon usage). The mutations to which all the remaining amino acids are subjected are performed in a more ad hoc manner. The only constraint that is imposed on these mutations is the requirement that the content of the amino acid whose repeats are investigated should be conserved. This is achieved by a minor bias of the replacements in the first and the second place in the codon. Within approximately 1000 mutational cycles, the sequence attains a steady composition with stable distribution of the repeats. The system becomes ready to perform the productive runs in which the sequence is receiving steady state modification. By analyzing the resulting sequences, the histogram of the runs of the codons of interest and the corresponding amino acids can be incremented.

After sufficient statistics are accumulated, one can compare the resulting histograms with the histograms obtained by the analysis of the natural sequences. The variation of the parameters that are involved in the simulation procedure enables one to tune the set of parameters to reproduce the properties of natural sequences. Because of the large number of free parameters, the optimization procedure may be quite tedious. Fortunately it is possible to conduct the process to exploit the separability of the parameter space, which enables one to determine some parameters independently of other parameters.

One way to exploit the separability property in the parameter space is to solve, before the simulation procedure, a partial problem in which only a limited set of parameters is involved. As an example, let us take the five histograms corresponding to a single amino acid that is coded by four codons. The four histograms representing the frequency of occurrence of runs of four codons can be treated as independent entities, which means that it is possible to run an independent numerical procedure devoted to each of the four individual histograms. The goal of the procedure is to reproduce the enriched part of the four histograms, disregarding the runs that conform to the Bernoulli model. The enriched part of the histogram can be obtained by extrapolating the values from above the crossover point between the "Bernoulli" and the "enrichment" region of the histogram to the Bernoulli region. The histogram obtained in this way is supposed to have a form proportional to $exp(-\alpha_c n)$ with the α_c parameter obtained from the fitting procedure for each codon. Within the numerical procedure, the histogram is receiving three types of alterations. The first operation is $h(2) \Rightarrow h(2) + 1$ (this operation adds new dinucleotides to the list of repeats, the corresponding probability is p_{2c}). The second operation is $h(n) \Rightarrow h(n) - 1$, $h(n \pm 1) \Rightarrow h(n \pm 1) + 1$ (elongation/shortening with probability $p_{e/s}$). The last operation corresponds to a point mutational event that splits the repeat with length n into two pieces with lengths k and $n\text{-}k\text{-}1$ with a probability $p_r + p_s$ and the corresponding histogram transformation reads $h(n) \Rightarrow h(n) - 1$, $h(k) \Rightarrow h(k) + 1$, $h(n\text{-}k\text{-}1) \Rightarrow h(n\text{-}k\text{-}1) + 1$. In the numerical procedure, one looks for such a combination of the parameters p_{2c}, $p_{e/s}$ and $p_r + p_{s,c}$ that conserves the form of the histogram during the series of the previously mentioned transformations. The method that is used for the determination of the parameters is a simple search on a two-dimensional network of points, which becomes denser in the course of the calculation. The number of transformations ranges from 10^3 (still far from the solution) to 10^6 (close to the solution of the problem). The drawback of the approach in which the mutational parameters are determined on the basis of the histograms of individual codons is due to the very poor statistics and large scatter in the histograms corresponding to natural sequences, which means that the α_c parameter is loosely defined. The histogram that displays the distributions of runs of the amino acid itself (this is not the sum of the histograms of the corresponding codons) shows better statistics, so that the conclusive phase of the mutational parameters determination is the simulation procedure on the model exon, as it is described above, in which the initial values of the p_{2c}, $p_{e/s}$, p_r, and $p_{s,c}$ parameters are taken from the results of the above described individual codon histogram transformation procedures.

Determination of the Percentage of Repeat Coverage

In random sequences, the probability of repeats is governed by the expression $N(n) = L(1\text{-}p)^2 p^n$, which means that on a logarithmic scale the histogram has the appearance of straight line with a (negative) slope $ln(p)$, where p is the probability of appearance of the element in question and L is the total length of the sequence. The histograms of natural sequences show the form of two superimposed exponentially decaying histograms, which can be fitted using two exponents. The two slopes on the logarithmic plot of the histogram correspond to the probabilities with which the particular amino acids are randomly distributed within the segments belonging to high and low density amino acid "isochores." It turns out that within the low density isochore, the rate of decay of the histogram $-\Delta(lnN[n])/\Delta[n])$ corresponds to the amino acid usage p_0 (sum of its codon usages). The rate of decay at large n values enables us to calculate p_r, which is the "density" of the corresponding amino acid in the repeat region. The abscissa value of the intersection n_{0r} of the two straight lines in the $ln(N[n])$ plot (Fig. 3) provides us with the information about the abundance of the repeat regions relative to the extent of the regions where the repeats are absent. At the point of intersection of the two straight lines one can make equal the two contributions to the histogram: $L_0(1\text{-}p_0)^2 exp(n_{0r} ln[p_0]) = L_r(1\text{-}p_r)^2 exp(n_{0r} ln[p_r])$. This equation can be solved for L_r/L_0, giving us

$$\frac{L_r}{L_0} = \frac{(1-p_0)^2}{(1-p_r)} \left[\frac{p_0}{p_r}\right]^{n_{0r}} \tag{M1}$$

Equation M1 connects the relative repeat coverage with the two slopes and their crossover point n_{0r}.

ACKNOWLEDGMENTS

The financial support of the Ministry of Education, Science

and Sport of the Republic of Slovenia is gratefully acknowledged. The anonymous referees are acknowledged for the suggestions on how to improve the manuscript and for bringing to our attention the works on the repeat enrichment modeling.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Bell, G.I. 1996. Evolution of simple sequence repeats. *Comput. Chem.* **20:** 41–48.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp B.A., and Wheeler, D.L. 2000. GenBank. *Nucleic Acids Res.* **28:** 15–18.

Borštnik, B., Pumpernik, D., and Lukman, D. 1993. Analysis of apparent 1/f spectrum in DNA sequences. *Europhys. Lett.* **23:** 389–394.

Borštnik, B., Pumpernik, D., Lukman, D., Ugarković, D., and Plohl, M. 1994. Tandemly repeated pentanucleotides in DNA sequences of eucaryotes. *Nucleic Acids Res.* **22:** 3412–3417.

Cox, R. and Mirkin, S.M. 1997. Characteristic enrichment of DNA repeats in different genomes. *Proc. Natl. Acad. Sci.* **94:** 5237–5242.

Dover, G. 2000. How genomic and developmental dynamics affect evolutionary processes. *Bioessays* **22:** 1153–1159.

Field, D. and Wills, C. 1998. Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutations pressures and variety of selective forces. *Proc. Natl. Acad. Sci.* **95:** 1647–1652.

Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Kruglyak, S., Durett, R.T., Schug, M.D., and Aquadro, C.F. 1998. Equilibrium distribution of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl. Acad. Sci.* **95:** 10774–10778.

Metzgar, D., Bytof, J., and Wills, C. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res.* **10:** 72–80.

Nakamura, Y., Gojobori, T., and Ikemura, T. 2000. Codon usage tabulated from international DNA sequence databases: Status for the year 2000. *Nucleic Acids Res.* **28:** 292.

Nishizawa, K., Nishizawa, M., and Kim, K.S. 1999. Amino acid and nucleotide recurrence in aligned sequences: Synonymous substitution patterns in association with global and local base compositions. *J. Mol. Biol.* **294:** 937–953.

Toth G., Gaspari, Z., and Jurka, J. 2000. Microsatellites in different eucaryotic genomes: Survey and analysis. *Genome Res.* **10:** 967–981.

Received August 7, 2001; accepted in revised form March 25, 2002.

Tandem Repeats in Protein Coding Regions of Primate Genes

Branko Borstnik and Danilo Pumpernik

Genome Res. 2002 12: 909-915
Access the most recent version at doi:10.1101/gr.138802