4056–4062 *Nucleic Acids Research, 1998, Vol. 26, No. 17*

Distinct frequency-distributions of homopolymeric DNA tracts in different genomes

Koen J. Dechering[+], Koen Cuelenaere[1], Ruud N. H. Konings and Jack A. M. Leunissen[1,*]

Department of Molecular Biology and [1]CAOS/CAMM Center, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

Received April 6, 1998; Revised and Accepted July 23, 1998

ABSTRACT

The unusual base composition of the genome of the human malaria parasite *Plasmodium falciparum* prompted us to systematically investigate the occurrence of homopolymeric DNA tracts in the *P.falciparum* genome and, for comparison, in the genomes of *Homo sapiens*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Escherichia coli* and *Mycobacterium tuberculosis.* Comparison of the observed frequencies with the frequencies as expected for random DNA revealed that homopolymeric (dA:dT) tracts occur well above chance in the eukaryotic genome. In the majority of these genomes, (dA:dT) tract overrepresentation proved to be an exponential function of the tract length. (dG:dC) tract overrepresentation was absent or less pronounced in both prokaryotic and eukaryotic genomes. On the basis of our results, we propose that homopolymeric (dA:dT) tracts are expanded via replication slippage. This slippage-mediated expansion does not operate on tracts with lengths below a critical threshold of 7–10 bp.

INTRODUCTION

The past decade has seen the initiation of a number of genome sequencing projects for organisms that are of interest as a model system or as a pathogen. An example of the latter category is the protozoan parasite *Plasmodium falciparum*, which is the main cause of malaria in man and responsible for two million deaths annually. With the aim of the development of new drugs or a vaccine, the biology of the parasite has been the subject of intensive study. One of the unique features of the parasite is the extraordinary base composition of its genome, first revealed by the sequencing of genes and intergenic regions, and confirmed later by data generated by the *P.falciparum* genome project. The overall A/T content of the parasite's genome is 81%, and can reach levels as high as 90% in non-coding regions (1). *Plasmodium falciparum* possesses the G/C poorest genome known so far (2).

Visual inspection of *P.falciparum* intergenic sequences reveals the extensive occurrence of long homopolymeric (dA:dT) stretches. These stretches are of interest as they have unique structural and functional properties. Crystallographic data have shown that homopolymeric (dA:dT) tracts adapt a rigid structure which is characterized by a high level of propeller twist and an increased base stacking. This allows the formation of additional non-Watson–Crick, bifurcated hydrogen bonds (3). Phasing of short (dA:dT) tracts within the helical repeat of normal *B*-DNA results in a macroscopic curvature of the DNA (4,5). It has been proposed that all general-sequence *B*-DNA gently writhes, with the net effect of all local bends being a straight helix. Introduction of a straight (dA:dT) tract distorts the array of compensating writhes and results in a curvature of the DNA (6). This curvature can play a role in the modulation of the transcriptional activity of genes (7), and enhance the affinity of the DNA for transcription factors such as the TATA-binding protein (8). Alternatively, (dA:dT) can modulate the access of transcription factors to the DNA via a local distortion of a nucleosome (9). In yeast, (dA:dT) tracts are functional promoter elements (10). Their effects are mediated by a modulation of the nucleosomal occupancy of the DNA rather than by the direct recruitment of *trans*-acting factors (11). Finally, homopolymeric (dA:dT) stretches are part of scaffold associated regions (SARs) that are supposed to anchor the chromatin loops in the nucleus (12,13). The SARs are also the place of residence of topoisomerase II, which controls the topology of DNA during replication, recombination and transcription. It has been proposed that the curvature induced by the homopolymeric (dA:dT) tracts in the SAR defines the sequence characteristics preferred by topoisomerase II (14,15).

Homopolymeric DNA tracts, or more generally, simple repetitive sequences, can give rise to slippage of the polymerase during replication. The internally repetitive DNA sequences allow the nascent strand to slip back or forward on the parental strand with one or more repeat units, resulting in an expansion or contraction of the new DNA strand (16). It has been proposed that slipped strand replication is a major force in the evolution of genes (17,18) and genomes (19) and it is supposed to be implicated in a large number of human genetic diseases (20,21). In addition to replication slippage, processes like unequal crossing over, mutation and selection affect the persistence of simple sequences. The distribution of simple sequences in the genome thus reflects an equilibrium between various mutational

*To whom correspondence should be addressed. Tel: +31 24 3652248; Fax: +31 24 3652977; Email: jackl@caos.kun.nl

[+]Present address: N. V. Organon, Target Discovery Unit, Room RH1204, PO Box 20, 5340 BH Oss, The Netherlands

This paper is dedicated to the late Ruud Konings

and selective forces (22). It is generally believed, however, that the initial variations in sequence lengths are provoked by replication slippage, and that this process provides the raw material upon which the other mechanisms act (16,23,24).

The functional and structural significance of homopolymeric (dA:dT) stretches together with initial observations that they are enriched in the genome of *P.falciparum* prompted us to investigate their occurrence in a systematic way. For comparison, we analyzed the occurrence of homopolymeric (dG:dC) stretches in the *P.falciparum* genome, and the occurrence of (dA:dT) and (dG:dC) tracts in the genomes of six other species (*Homo sapiens*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Escherichia coli* and *Mycobacterium tuberculosis*), which are widely ranged across the evolutionary spectrum.

MATERIALS AND METHODS

Analysis of *P.falciparum* genomic sequences was performed on 17 contigs (#7289, #7290, #7292, #7294, #7296, #7297, #7299, #7300, #7302, #7316, #7327, #7355, #7404, #7455, #7535, #7623, #7651) obtained from the Institute for Genomic Research (ftp://ftp.tigr.org), and compiled from the sequence data of chromosome 2 of strain 3D7 of *P.falciparum*. For the analysis of the human genome, 10 contigs (BK992D9, DJ121G13, DJ211D12, DJ30P20, DJ431A14, DJ106H8, DJ170A21, DJ272J12, DJ389A20, DJ79C4) from sequence data of chromosomes 1, 6, 20, 22 and X were obtained from the Sanger Center (ftp://ftp.sanger.ac.uk). Sequences of *C.elegans* and *A.thaliana* were obtained from the EMBL nucleotide library (*C.elegans*: CEY105C5, CEY106G6; *A.Thaliana*: ATFCA0; ATFCA1; ATFCA2; ATFCA3; ATFCA4; ATFCA5; ATFCA6; ATFCA7; ATFCA8). Analysis of the *S.cerevisiae*, *E.coli* K12 and *M.tuberculosis* H37Rv genomes was on the completed genomic sequences obtained from ftp://genome-ftp.stanford.edu; ftp://ftp.genetics.wisc.edu; and ftp://ftp.sanger.ac.uk, respectively.

To investigate the occurrence of homopolymeric (dA:dT) stretches in different *P.falciparum* genome regions, sequences with well-defined features with respect to the organization into intron, exon and flanking sequences were selected from the EMBL database, release 52. Sequences encoding structural RNA or originating from plastids or mitochondria were omitted. A total number of 241 unique sequences were selected with a total length of 608 kb. Using a small Tcl script, the sequences were dissected in gene-flanking, coding and intron according to the tables of features accompanying the sequences.

The basic characteristics of all sequences analyzed are presented in Table 1. Sequence analysis was performed on a Silicon Graphics Challenge running Irix 5.3 using the GCG package (25) version 8.1. The expected frequency of finding a homopolymeric non-overlapping (dA:dT) tract of length N in either orientation was calculated assuming a zero-order Markov chain as described previously (26):

$$f_N^{\exp} = (f_{A_{N=1}}^{obs})^N(1 - f_{A_{N=1}}^{obs})^2 + (f_{T_{N=1}}^{obs})^N(1 - f_{T_{N=1}}^{obs})^2$$

Note that this equation calculates the frequency of a non-overlapping (dA:dT) tract and takes into account the frequencies of the two adjacent nucleotides.

To assess whether the overrepresentation of homopolymeric (dA:dT) tracts is due to compositional inhomogeneities in the genome, regression analysis was carried out between the local A/T content of the contigs and the frequency of occurrence of overrepresented homopolymeric (dA:dT) tracts. To this end, the abundance of homopolymeric (dA:dT) tracts ≥ 10 bp and the A/T content were determined in a window of 1000 bp that was shifted along the sequence with a 950 bp interval. Regression analysis was performed using the regression module of Microsoft Excel 97.

RESULTS

The *P.falciparum* genome is enriched for short (dG:dC) tracts and long (dA:dT) tracts

The malaria genome project, which was established in 1996 and ultimately aims at sequencing all the 2.5×10^7 nt of the genome, is in full progress and has already provided a wealth of sequence information (27). As chromosome 2 was the first chromosome for which a complete contig map was established in yeast artificial chromosomes (28), most progress has been made in sequencing this chromosome. To date, this has resulted in 21 807 individual sequence reads that can be assembled into 17 contigs that cover 967 kb of chromosome 2. As the estimated size of chromosome 2 is 1.03 Mb (28), the contigs encompass 94% of the chromosome and can be considered representative for its DNA sequence. The overall A/T content of the chromosome 2 contigs is 80% (Table 1), which corresponds well to the numbers that have been reported previously for the *P.falciparum* genome (29,30). We determined the numbers of non-overlapping homopolymeric (dA:dT) and (dG:dC) tracts present in either orientation in the chromosome 2 contigs. Table 2 shows the numbers of tracts for $2 \leq N \leq 10$. The results show that all (dG:dC) tracts appear at higher frequencies than is expected on basis of a random distribution of nucleotides. However, (dG:dC) tracts longer than 9 nt are not observed. (dA:dT) tracts of 2 or 3 nt are slightly overrepresented whilst tracts of 4 nt show a minor underrepresentation. (dA:dT) tracts with lengths of 5–9 nt are overrepresented, but to a lesser extent than (dG:dC) tracts of similar lengths. Standard χ^2 analysis revealed that in all cases the deviations of the observed frequencies from the expected frequencies are statistically significant at $P < 0.001$ (not shown).

Whereas (dA:dT) tracts <10 nt appear at frequencies that are close to expectation in the *P.falciparum* genome, the occurrence of tracts >10 bp deviates strongly from expectation. Figure 1A shows the relative frequency of occurrence as a function of the tract length for the non-overlapping homopolymeric DNA tracts found in the chromosome 2 contigs. Interestingly, the frequency distribution of (dA:dT) tracts >12 bp can be fit by a single semi-logarithmic function that exhibits a far greater dependence on N than the function that describes the expected frequencies. (dA:dT) tracts with lengths of up to 47 bp are observed, which would be very unlikely to occur at a random distribution of nucleotides. The contribution of homopolymeric (dA:dT) tracts to the genome is considerable, as the sum of the lengths of all stretches >7 bp is 44 506 bp, which accounts for nearly 5% of the sequences encompassed by the chromosome 2 contigs. In conclusion, the *P.falciparum* genome as represented by the chromosome 2 contigs is highly enriched for short (dG:dC) tracts and for long (dA:dT) tracts.

We assessed whether the observed overrepresentation of homopolymeric (dA:dT) tracts is caused by compositional inhomogeneities in the *P.falciparum* genome. To this end, we determined the base composition and frequency of occurrence of homopolymeric (dA:dT) tracts of ≥ 10 bp in windows of 1000 bp. Regression analysis revealed a very weak correlation between the

Table 1. Summary and basic characteristics of the sequence data analyzed in this study

Organism	Genome size	number of contigs analyzed	number of base pairs analyzed	A/T content (%)	r^2 value*
P. falciparum	$2.5x10^7$	17	967,532	80.0	0.170
H. sapiens	$3.4x10^9$	10	1,711,657	57.1	0.032
S. cerevisiae	$1.3x10^7$	16	13,119,303	61.7	0.004
C. elegans	$8.0x10^7$	2	1,482,646	66.7	0.000
A. thaliana	$1.0x10^8$	9	1,840,460	64.1	0.004
M. tuberculosis	$4.4x10^6$	1	4,411,522	34.4	n.d.
E. coli	$4.6x10^6$	1	4,639,221	49.2	n.d.
***P. falciparum* genome compartment**	**number of sequences analyzed**	**number of base pairs analyzed**	**A/T content (%)**		
gene-flanking	228	123,193	82.6		
coding	225	356,382	71,1		
intron	118	24,055	86,6		

The estimated sizes of the genomes of *P.falciparum*, *H.sapiens*, *C.elegans* and *A.thaliana* were taken from (43), (44), (45) and (46) respectively. The A/T content was determined from the sequence data. *r^2 values obtained by linear regression analysis between the A/T content and the frequency of occurrence of homopolymeric (dA:dT) tracts $\leq$10bp in windows of 1000 bp.

Table 2. Occurrence of homopolymeric tracts in *P.falciparum* chromosome 2

(dA:dT)					(dG:dC)			
tract length	number of tracts	observed relative frequency	expected relative frequency	obs/exp	number of tracts	observed relative frequency	expected relative frequency	obs/exp
2	158,165	0.1635	0.1152	1.4	21,702	0.0224	0.0162	1.5
3	68,736	0.0710	0.0461	1.5	2,903	0.0030	0.0016	2.4
4	16,658	0.0172	0.0184	0.9	510	0.0005	0.0002	5.1
5	7,856	0.0081	0.0074	1.1	129	0.0001	$1.7\,10^{-5}$	15.0
6	3,522	0.0036	0.0030	1.2	35	$3.6\,10^{-5}$	$1.7\,10^{-6}$	53.6
7	1,766	0.0018	0.0012	1.5	12	$1.2\,10^{-5}$	$1.7\,10^{-7}$	228.2
8	976	0.0010	0.0005	2.1	9	$9.3\,10^{-5}$	$1.8\,10^{-8}$	971.4
9	544	0.0006	0.0002	3.0	6	$6.2\,10^{-6}$	$1.9\,10^{-9}$	2691.3
10	358	0.0004	0.0001	4.9	0	0	$1.9\,10^{-10}$	0

The table lists the number of non-overlapping homopolymeric (dA:dT) and (dG:dC) tracts of lengths N for $2 \leq N \leq 10$ as found in 17 contigs spanning 967 532 nt of chromosome 2. The observed and expected relative frequencies of occurrence are given together with the ratio of the observed to the expected frequencies (overrepresentation).

A/T content in the window and the abundance of homopolymeric tracts (Table 1). Thus, the abundance of homopolymeric tracts is independent of local inhomogeneities in the genome.

Enrichment for (dA:dT) tracts is restricted to non-coding DNA

As coding and non-coding regions are subject to different functional constraints, it is likely that the occurrence of homopolymeric DNA tracts will vary between these regions. Therefore, we analyzed the occurrence of homopolymeric stretches in the different regions of the *P.falciparum* genome. 241 sequences were selected from the EMBL database and the number of tracts in the coding, intron and gene-flanking regions of the *P.falciparum* genome was scored. Table 1 summarizes the basic features of the data we have analyzed. The coding regions have an A/T content of 71% whereas the A/T content reaches 81% in the gene-flanking regions and 87% in the introns.

Figure 1B–D shows the length dependent occurrence of homopolymeric stretches in the different genome regions. Short (dG:dC) tracts show a minor overrepresentation in the introns while higher levels of overrepresentation are seen in the gene-flanking and coding sequences. (dG:dC) runs >9 bp are absent. In accordance with the analysis of the chromosome 2 contigs, short (dA:dT) tracts of N <10 bp appear at frequencies close to expectation in all regions. However, dramatic differences between the genomic regions become apparent for (dA:dT) tracts >10 bp. Whereas the coding regions are limitedly enriched for (dA:dT) tracts >10 nt, these tracts appear at frequencies well above chance in the non-coding regions. In the latter regions, the frequency distributions of (dA:dT) tracts show a characteristic bipartite pattern very similar to that observed for the chromosome 2 contigs. These

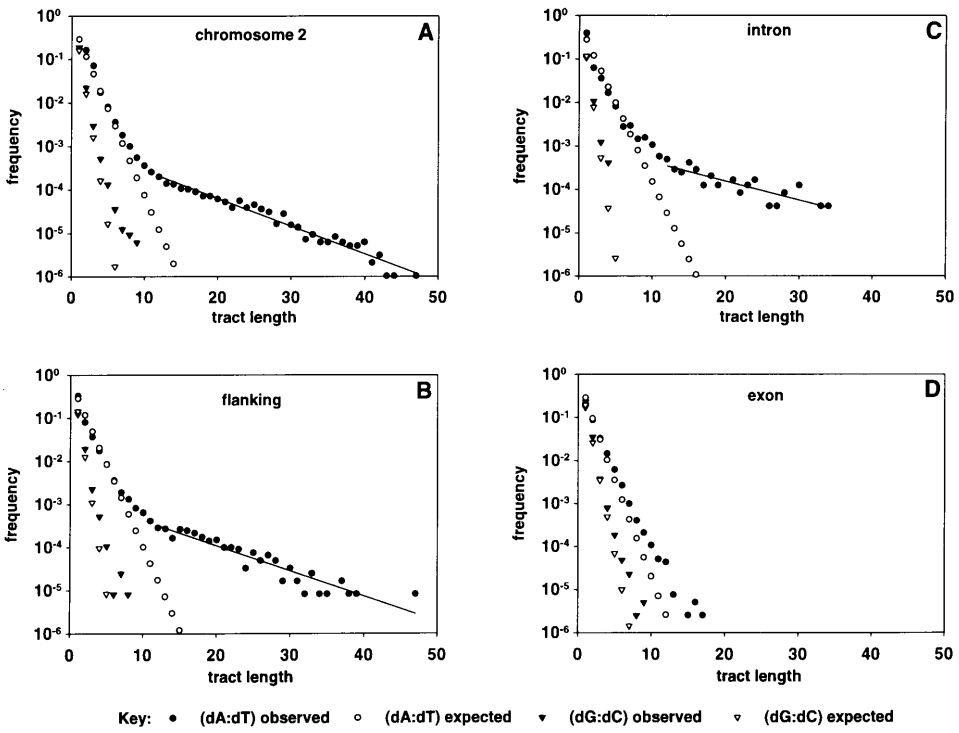

Figure 1. Frequency distributions of homopolymeric tracts in the *P.falciparum* genome. The figure shows the frequency distributions of homopolymeric runs as observed in chromosomal (**A**), gene-flanking (**B**), intron (**C**) and protein encoding DNA (**D**) of *P.falciparum*.

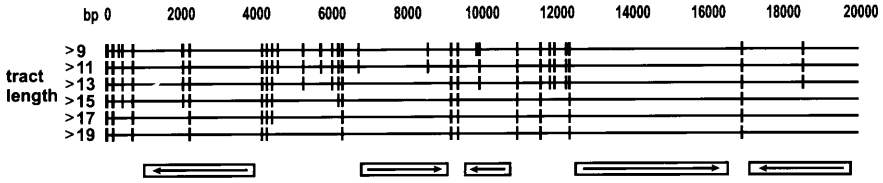

Figure 2. Homopolymeric (dA:dT) tracts cluster in the non-coding regions of the *P.falciparum* genome. The figure shows the distribution of homopolymeric (dA:dT) tracts (vertical bars) along a representative 20 kb region of chromosome 2, together with a prediction of open reading frames (open boxes). Arrows indicate the directions of the open reading frames.

results indicate that the enrichment for long (dA:dT) tracts as seen in the analysis of the chromosome 2 contigs can be largely attributed to the gene-flanking and intron sequences. To support this notion, we plotted the regions occupied by homopolymeric tracts together with a prediction of open reading frames along a 20 000 bp region of chromosome 2 (Fig. 2). In this representation it can be seen that long homopolymeric (dA:dT) tracts and open reading frames indeed appear in a mutually exclusive pattern.

(dA:dT) tract enrichment is a general eukaryotic phenomenon

At a random distribution of nucleotides, homopolymeric (dA:dT) stretches would occur relatively frequently in an A/T rich genome whereas (dG:dC) tracts may be virtually absent. In the *P.falciparum* genome, for instance, (dA:dT) tracts of 8 bp are expected to occur once every 2000 nt whereas (dG:dC) tracts of similar length are expected to occur only once every 5.6×10^7 nt (Table 2). It is conceivable that the overrepresentation of (dA:dT) tracts in the *P.falciparum* genome is provoked by the intrinsic high frequency of randomly occurring tracts, which may serve as the substrate for slippage-mediated expansion. If such a process would operate with equal efficiency on A/T rich and G/C rich DNA, it would lead to (dA:dT) tract enrichment in an A/T rich genome, and (dG:dC) tract enrichment in a G/C rich genome. In this view, in a genome with an A/T to G/C ratio of 1, homopolymeric (dG:dC) tract overrepresentation should equal (dA:dT) tract overrepresentation. To address this hypothesis, we analyzed the occurrence of homopolymeric tracts in several genomes with varying A/T contents.

In contrast to the prediction, in none of the genomes analyzed is a high overrepresentation of (dG:dC) tracts observed (Fig. 3).

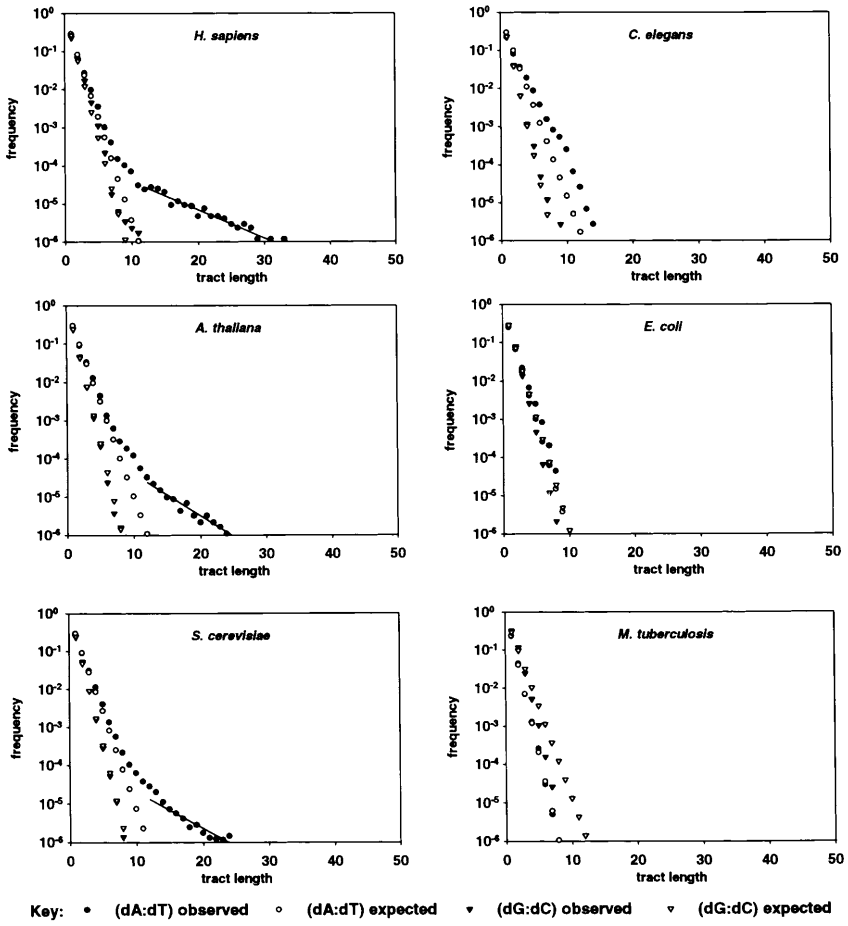

Figure 3. Length dependent occurrence of homopolymeric tracts in different eukaryotic and prokaryotic genomes.

Instead, the occurrence of (dG:dC) tracts is close to expectation (*S.cerevisiae*, *A.thaliana*), shows an underrepresentation (*E.coli*, *M.tuberculosis*) or a relatively minor overrepresentation (*P.falciparum*, *H.sapiens*, *C.elegans*). These results show that a higher G/C content does not lead to a dramatic increase in the overrepresentation of (dG:dC) tracts, and might indicate that replication slippage operates less efficiently on G/C rich DNA.

The analysis of the various genomes furthermore shows that the patterns of occurrence of (dA:dT) tracts are clearly distinct between prokaryotes and eukaryotes (Fig. 3). In the two prokaryotes we have analyzed, (dA:dT) tracts appear at frequencies close to expectation. In eukaryotes, poly(dA:dT) tracts are generally overrepresented following a characteristic bipartite pattern. The frequency distribution of (dA:dT) tracts in the genomes of *P.falciparum*, *H.sapiens*, *S.cerevisiae* and *A.thaliana* can be fitted by two exponential functions that break in the 8–12 bp region. A strikingly divergent pattern of (dA:dT) tract overrepresentation is provided by *C.elegans*. In this organism, the curve that fits the observed frequencies shows a slight bulge in the 8–10 bp region, but then continues parallel to the curve that represents the distribution of the expected frequencies. Furthermore, (dA:dT) tracts >14 bp were not observed in the *C.elegans* genome whereas in all other eukaryotes tracts reach lengths of over 25 bp.

For all eukaryotic genomes, regression analysis between the local A/T content and the density of overrepresented homopolymeric (dA:dT) tracts revealed that these are not correlated (Table 1). This indicates that the overrepresentation of homopolymeric (dA:dT) tracts in the genomes of higher eukaryotes is not due to the presence of A/T rich compartments, or isochores.

Overrepresentation of (dA:dT) tracts is an exponential function of the tract length

In contrast to the situation in *C.elegans*, where tracts >10 bp appear at frequencies that are at a steady 15-fold above expectation, the overrepresentation of (dA:dT) tracts in the genomes of the other eukaryotes is an exponential function of the tract length. This can best be seen in Figure 4, where the ratio of the observed to the expected frequency is plotted against the tract length. In the genomes of *P.falciparum*, *H.sapiens*, *S.cerevisiae* and *A.thaliana*, longer tracts are more strongly overrepresented then shorter tracts. The overrepresentation of tracts >10 bp can be fitted by a single exponential function that depends on the tract length. Interestingly, these functions are very similar for the different genomes, suggesting a shared mechanism for the accumulation and maintenance of (dA:dT) tracts.

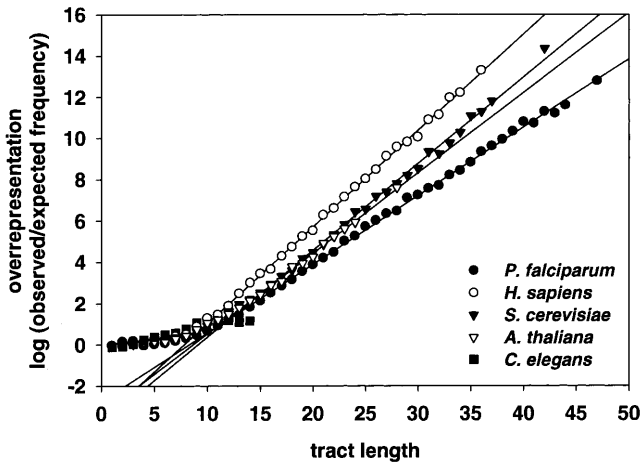

Figure 4. Overrepresentation of (dA:dT) tracts is an exponential function of the tract length in the genomes of *P.falciparum*, *H.sapiens*, *S.cerevisiae* and *A.thaliana*, but not in the *C.elegans* genome. Overrepresentation of (dA:dT) tracts, expressed as the ratio between the observed and the expected frequency of occurrence, is plotted against the tract length.

DISCUSSION

The data presented here show that the eukaryotic genome is enriched for homopolymeric (dA:dT) tracts. With the exception of *C.elegans*, the occurrence of (dA:dT) tracts shows a bipartite pattern that can be described by two exponential functions. First, short tracts of $2 \leq N \leq 7$ occur at frequencies that are close to the predicted values. Second, tracts of $N > 10$ show a length dependent overrepresentation and can reach lengths of >30 nt that are up to 10^{12}-fold overrepresented. By contrast, (dG:dC) tracts are not or only weakly overrepresented. In the few instances that enrichment for (dG:dC) tracts is seen, the overrepresentation cannot be described by a simple exponential function (not shown) and never exceeds 10^4-fold over chance. The detailed analysis of the *P.falciparum* genome shows that overrepresentation of (dA:dT) tracts is largely restricted to non-coding DNA.

Slipped strand mispairing during replication rather than unequal crossing-over is seen as the major force in the generation of length variation of simple sequence repeats (16,23), and this process can also account for the variation in length of homopolymeric runs, which are the simplest forms of simple sequence repeats (19,31). The driving force in the expansion of homopolymeric tracts might originate from a biased action of the slippage process or from specific retention of expanded tracts (16). In either case, the process has a self-accelerating component, as expanded tracts increase the likelihood for additional slippage events, which in turn lead to additional expansion.

Two observations from the data presented here stand out. First, the overrepresentation of (dA:dT) tracts >10 bp in the genomes of *P.falciparum*, *H.sapiens*, *S.cerevisiae* and *A.thaliana* is an exponential function of the tract length. Such a distribution is consistent with models in which replication slippage is responsible for the expansion of homopolymeric DNA tracts (22,32). Interestingly, tracts <7 bp appear at frequencies close to expectation, indicating that they are immune to slippage mediated expansion. This suggests that there is a critical threshold that determines whether a homopolymeric tract can be subjected to slippage-mediated expansion. A length <7 bp is below the threshold for expansion. By contrast, the lengths of (dA:dT) tracts >10 bp are above the threshold and these tracts accumulate in the genome as a result of slippage-mediated expansion. A threshold value identical to that observed here can be determined from the data presented in a study of the length-dependent occurrence of homopolymeric (dA:dT) tracts in the *Dictyostelium discoideum* genome (26), and from data on the length dependent occurrence of DNA repeats in a variety of genomes (33). We conclude, therefore, that irrespective of the organism and of the nature of the repeat element, the minimum length requirement for a simple sequence repeat to undergo expansion by replication slippage is 7–10 bp.

Our data indicate that for the majority of the eukaryotic genomes, the expansion of (dA:dT) tracts >10 bp can be described by a single exponential function, which is very similar for the different genomes. A striking exception is provided by *C.elegans* where the frequency distribution of (dA:dT) tracts is clearly distinct from that of the other eukaryotes. Although the curve that fits the length dependent occurrence of (dA:dT) tracts in the *C.elegans* genome does change its slope slightly in the 8–10 bp region, it differs drastically from the curves seen for the other eukaryotes. The reason for this is unclear. The overall level of sequence simplicity in *C.elegans* is similar to that seen in other eukaryotes (19), indicating that the mechanisms responsible for the generation and maintenance of simple sequence repeats operate with comparable efficiency in *C.elegans*. Therefore, the distinct pattern of (dA:dT) tract overrepresentation most probably results from distinct selective forces. The nature of these forces remains unresolved.

A second important observation made here is that the frequency distribution of (dG:dC) tracts is very different between the various eukaryotic genomes. Some genomes are enriched for (dG:dC) tracts, whereas other genomes exhibit an underrepresentation. The A/T rich *P.falciparum* genome is enriched for (dG:dC) stretches ≤ 9 bp. As the lengths of the vast majority of these stretches are below the threshold for slippage, expansion of (dG:dC) tracts by replication slippage is precluded, and the overrepresentation of short (dG:dC) tracts most probably has evolved by other mechanisms. The genomes of the other eukaryotes studied here are more G/C rich, and will, by chance, have higher densities of (dG:dC) tracts. These stretches provide the substrate for slippage-induced expansion. Yet, overrepresentation of (dG:dC) tracts is absent or, in cases where it is observed, does not reach the high level seen for (dA:dT) tracts. This indicates that (dG:dC) tracts are less efficiently expanded by slipped strand replication. This might not be surprising: slippage during replication requires the local melting of a DNA duplex, and the greater stability of (dG:dC) duplexes in comparison to (dA:dT) duplexes might prevent slippage of polymeric (dG:dC) tracts. Accordingly, it has been shown that the efficiency of *in vitro* slippage synthesis of simple sequence DNA using short primers is dependent on the A/T content of the primers. Whereas A/T rich primers mediate slippage synthesis at a high rate, primers consisting purely of G/C nucleotides poorly generate simple sequence repeats (34). Thus, the low enrichment for (dG:dC) tracts in the eukaryotic genome most likely indicates that they are less efficiently expanded by slippage-like events.

Superimposed on the results of slipped strand replication are the actions of unequal crossing-over, mutation, gene conversion and selection that all act on the persistence of simple sequence

DNA (22). As coding regions are subject to strong selection, the ways in which slippage-derived sequences accumulate in them are more restricted than they are in non-coding regions (19). Accordingly, we observed that homopolymeric tracts are less strongly overrepresented in the *P.falciparum* coding than in the non-coding regions. Furthermore, overrepresentation of homopolymeric tracts is absent in prokaryotes. This is consistent with the view that prokaryotes possess a streamlined genome, which allows for rapid replication and cell division (16,35). It is not clear whether the homopolymeric (dA:dT) tracts in the non-coding regions of the eukaryotic genome are also under selective pressure or represent junk DNA. Such junk, or 'selfish' DNA, is evolutionary neutral and does not affect the phenotype of its host (36). As selfish DNA is not under control of selective forces, it can only accumulate by virtue of a biased action of the replication machinery or by an ability to self-replicate as, for instance, is seen for transposons. Replication slippage itself might not be biased towards the generation of duplications. In experimental contexts, simple sequence repeats subjected to slippage events are unstable and acquire deletions rather than insertions (37–39). Thus, selective rather than stochastic principles might underlie the overrepresentation of homopolymeric runs. Selective advantages of homopolymeric (dA:dT) runs might originate from their structure forming abilities. Since one of the canonical features of the structure-forming ability of a DNA sequence is its length-dependence (40), any selective advantage given by the structure-forming ability of a (dA:dT) tract should be reflected in a length-dependent enrichment (33). Our results show that in the *P.falciparum*, *H.sapiens*, *S.cerevisiae* and *A.thaliana* genomes, tracts >10 nt show an overrepresentation that is an exponential function of the tract length. Longer tracts are up to a billion-fold overrepresented whereas short tracts only show a minor overrepresentation. This strongly suggests that the structure-forming abilities of (dA:dT) tracts offer selective advantages that lead to their overrepresentation. The precise nature of this selective advantage remains unclear but might originate from a functional role of (dA:dT) tracts in the modulation of transcription and/or in the organization of the chromatin structure (9,11,14,41).

All functional roles of homopolymeric (dA:dT) tracts reported in the literature relate to the structural organization of chromatin in the nucleus. Would that explain their overrepresentation in eukaryotes, or are they just byproducts of DNA metabolism in eukaryotic cells and represent selfish DNA? Regardless of a possible functional role, their presence will have an impact on the structure and organization of the DNA. Our analysis shows that (dA:dT) tracts make up a considerable 5% of the *P.falciparum* genome. In this organism, (dA:dT) tracts have been implicated in intrachromosomal recombination events that contribute to antigenic variation (42). Although the exact nature of the principles that lead to an overrepresentation of (dA:dT) tracts may be hard to resolve, having long homopolymeric stretches in the genome obviously has important biological consequences.

ACKNOWLEDGEMENTS

The authors wish to thank Harm Nijveen for discussion and computer programming. Nicolette Lubsen and Henk Stunnenberg are gratefully acknowledged for critical review of the manuscript.

REFERENCES

1 Hyde,J.E. and Sims,P.F.G. (1987) *Gene*, **61**, 177–187.
2 Musto,H., Rodriguez Maseda,H. and Bernardi,G. (1995) *Gene*, **152**, 127–132.
3 Nelson,H.C.M., Finch,J.T., Luisi,B.F. and Klug,A. (1987) *Nature*, **330**, 221–226.
4 Marini,J.C., Levene,S.D., Crothers,D.M. and Englund,P.T. (1982) *Proc. Natl Acad. Sci. USA*, **79**, 7664–7668.
5 Koo,H.-S., Wu,H.-M. and Crothers,D.M. (1986) *Nature*, **320**, 501–506.
6 Goodsell,D.S., Kaczor-Grzeskowiak,M. and Dickerson,R.E. (1994) *J. Mol. Biol.*, **239**, 79–96.
7 Cress,D.W. and Nevins,J.R. (1996) *Mol. Cell. Biol.*, **16**, 2119–2127.
8 Parvin,J.D., McCormick,R.J., Sharp,P.A. and Fisher,D.E. (1995) *Nature*, **373**, 724–727.
9 Zhu,Z. and Thiele,D.J. (1996) *Cell*, **87**, 459–470.
10 Struhl,K. (1985) *Proc. Natl Acad. Sci. USA*, **82**, 8419–8423.
11 Iyer,V. and Struhl,K. (1995) *EMBO J.*, **14**, 2570–2579.
12 Kaes,E., Izaurralde,E. and Laemmli,U.K. (1989) *J. Mol. Biol.*, **210**, 587–599.
13 Gasser,S.M. and Laemlli,U.K. (1986) *Cell*, **46**, 521–530.
14 Kaes,E. and Laemmli,U.K. (1992) *EMBO J.*, **11**, 705–716.
15 Miassod,R., Razin,S.V. and Hancock,R. (1997) *Nucleic Acids Res.*, **25**, 2041–2046.
16 Levinson,G. and Gutman,G.A. (1987) *Mol. Biol. Evol.*, **4**, 203–221.
17 Treier,M., Pfeifle,C. and Tautz,D. (1989) *EMBO J.*, **8**, 1517–1525.
18 Hancock,J.M. (1993) *Nucleic Acids Res*, **21**, 2823–2830.
19 Hancock,J.M. (1995) *J. Mol. Evol.*, **41**, 1038–1047.
20 Richards,R.I. and Sutherland,G.R. (1994) *Nature Genet.*, **6**, 114–116.
21 Bates,G. and Lehrach,H. (1994) *Bioessays*, **16**, 277–284.
22 Walsh,J.B. (1987) *Genetics*, **115**, 553–567.
23 Tautz,D., Trick,M. and Dover,G.A. (1986) *Nature*, **322**, 652–656.
24 Hancock,J.M. (1996) *Bioessays*, **18**, 421–425.
25 Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.
26 Marx,K.A., Hess,S.T. and Blake,R.D. (1993) *J. Biomol. Struct. Dynamics*, **11**, 57–66.
27 Dame,J.B., Arnot,D.E., Bourke,P.F., Chakrabarti,D., Christodoulou,Z., Coppel,R.L., Cowman,A.F., Craig,A.G., Fischer,K., Foster,J. *et al.* (1996) *Mol. Biochem. Parasitol.*, **79**, 1–12.
28 Lanzer,M., de Bruin,D. and Ravetch,J.V. (1993) *Nature*, **361**, 654–657.
29 McCutchan,T.F., Dame,J.B., Miller,L.H. and Barnwell,J. (1984) *Science*, **225**, 808–811.
30 Pollack,Y., Katzen,A.L., Spira,D.T. and Golenser,J. (1982) *Nucleic Acids Res.*, **10**, 539–546.
31 Newfeld,S.J., Tachida,H. and Yedvobnick,B. (1994) *J. Mol. Evol.*, **38**, 637–641.
32 Tachida,H. and Iizuka,M. (1992) *Genetics*, **131**, 471–478.
33 Cox,R. and Mirkin,S.M. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 5237–5242.
34 Schlötterer,C. and Tautz,D. (1992) *Nucleic Acids Res.*, **20**, 211–215.
35 Blattner,F.R., Plunket,III,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Wayne Davis,N., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) *Science*, **277**, 1453–1462.
36 Doolittle,W.F. and Sapienza,C. (1980) *Nature*, **284**, 601–603.
37 Henderson,S.T. and Petes,T.D. (1992) *Mol. Cell. Biol.*, **12**, 2749–2757.
38 Tran,H.T., Degtyareva,N.P., Koloteva,N.N., Sugino,A., Masumoto,H., Gordenin,D.A. and Resnick,M.A. (1995) *Mol. Cell. Biol.*, **15**, 5607–5617.
39 Farber,R.A., Petes,T.D., Dominska,M., Hudgens,S.S. and Liskay,R.M. (1994) *Hum. Mol. Genet.*, **3**, 253–256.
40 Vologodskii,A. (1992) *Topology and Physics of Circular DNA*. CRC Press, Boca Raton, FL.
41 Saitoh,N., Goldberg,I. and Earnshaw,W.C. (1995) *Bioessays*, **17**, 759–766.
42 Pologe,L.G., De Bruin,D. and Ravetch,J.V. (1990) *Mol. Cell. Biol.*, **10**, 3243–3246.
43 Weber,J.L. (1988) *Exp. Parasitol.*, **66**, 143–170.
44 Cavalier-Smith,T. (1985) In Cavalier-Smith,T. (ed.), *The Evolution of Genome Size*. John Wiley, New York, pp. 69–103.
45 Hodgkin,J., Plasterk,R.H.A. and Waterson,R.H. (1995) *Science*, **270**, 410–414.
46 Schmidt,R., West,J., Love,K., Lenehan,Z., Lister,C., Thompson,H., Bouchez,D. and Dean,C. (1995) *Science*, **270**, 480–483.