CROATICA CHEMICA ACTA
CCACAA **77** (1–2) 73–81 (2004)

ISSN-0011-1643
CCA-2902
Original Scientific Paper

Spectral Densities and Frequencies in the Power Spectrum of Higher Order Repeat Alpha Satellite in Human DNA Molecule*

Vladimir Paar,[a], Nenad Pavin,[a] Ivan Basar,[a] Marija Rosandić,[b] Ivica Luketin,[c] and Sonja Durajlija Žinić[d]**

[a]*Department of Physics, Faculty of Science, University of Zagreb, Zagreb, Croatia*

[b]*Department of Internal Medicine, University Hospital Rebro, Zagreb, Croatia*

[c]*Department of Physics, Faculty of Science, University of Split, Split, Croatia*

[d]*Ruđer Bošković Institute, Zagreb, Croatia*

RECEIVED JANUARY 9, 2003; REVISED APRIL 22, 2003; ACCEPTED APRIL 25, 2003

Fast Fourier transform was applied to the central segment of a fully sequenced genomic segment from the centromeric region in human chromosome 7 (GenBank/AC017075.8, 193277 bp), which is characterized by alpha satellite higher order repeats (HOR). Frequencies and spectral densities were computed for all prominent peaks in the Fourier spectrum. We have additionally introduced a peak to noise ratio as effective spectral density in order to account for frequency variations of the noise level. We have shown that a very good description of computed Fourier frequencies can be obtained by using the multiple formula with the fundamental frequency corresponding to the 2734-bp HOR sequence. The peak at f_{16} corresponds to the 171-bp monomer. Above the frequency f_{16}, the most pronounced peaks are mostly at multiples of f_{16} (monomer-multiples). The lowest sixteen monomer-multiples kf_{16} are locally dominant in spectral densities. The first monomer-multiple that is not locally dominant in spectral density is at $k = 17$. Above $k = 27$, the maximum of spectral density is systematically shifted to several neighboring higher frequency multiples. On the basis of the Fourier spectrum, the 171-bp monomer unit was subdivided into three approximately 57-bp subrepeats, which were further subdivided into 12-bp, 14-bp and 17-bp basic subrepeats.

Key words
human DNA
alpha satellite
higher order repeat
Fourier spectrum

INTRODUCTION

The centromeric regions of human and other primate chromosomes contain tandemly repeated DNA and the alpha satellite is the most extensively studied one.[1–7] Alpha satellite DNA is arranged in tandem arrays, in which the monomer subunits are approximately 171 bp in length. They can be further organized into highly homologous multimeric higher-order repeats (HOR), which can give a characteristic periodicity to each tandem array.[8–11]

* Dedicated to Professor Nenad Trinajstić on the occasion of his 65th birthday.

** Author to whom correspondence should be addressed. (E-mail: paar@hazu.hr)

Previously, a 16-mer HOR consisting of ten copies was identified in chromosome 7 (Refs. 10,11) and using the key-string algorithm (KSA), we have recently identified 54 HOR copies[12] using the 193277-bp complete human genomic sequence AC017075.8 (Ref. 13) from the centromeric region of human chromosome 7.

Statistical analysis of genomic sequences using the Fourier spectral analysis has been mostly applied to the studies of exons and introns.[14–25] Other statistical approaches included the enhance algorithm for distance frequency distribution,[26] random walk analysis,[27–30] chaos game representation,[31] wavelet transform,[32] advanced computer algorithm to identify approximate periodic repetitions up to 40 bp in length,[33] Shannon information analysis,[34] portrait method,[35] spectral approach,[36] segmentation algorithm based on entropic divergence,[37] *etc*. Genetic sequence data banks were scanned to analyze trinucleotide and pentanucleotide repeats.[38,39]

Besides being used to study long-range correlations ($1/f_\beta$ – behavior), the Fourier analysis, which can identify repeats of certain segments of the same length in nucleotide sequences, was applied to search for hidden periodicities in DNA sequences. A search for periodic regularities with periods from two to ten base pairs, carried out on a sample set of human exons and introns, showed a pronounced peak of period three.[16,20,21,30,38] For some gene sequences, several different periodicities could be observed in the power spectrum. For example, prominent peaks corresponding to the periods of 2 bp, 17 bp and 93–98 bp have been found in the power spectra for nucleotide distribution around the replication origin site of *E. coli*.[40] In the power spectrum of beta-globulin gene sequences, prominent peaks were found at 3 bp, 10 bp, 11.2 bp, 21.3 bp, 106.4 bp and 204.8 bp.[41] In some cases, the real genome was compared with its white noise genome, corresponding to the random sequence based on frequencies of four kinds of nucleotides appearing in the real genome.[35]

In this paper, we investigate spectral densities up to high frequencies in the power spectrum of the central (HOR) domain of AC017075.8.

SPECTRAL DENSITIES OF HOR ALPHA SATELLITE DNA

In our previous analyses of complete nucleotide sequence, we found that the large central domain of clone AC017075.8, from positions 31209 to 179354, exhibits a highly organized super-repeat pattern.[12] This central domain represents 76 % of the total length of AC017075.8 and 54 copies of the 16-mer (2734-bp HOR), which are highly convergent (mutual divergence of less than 0.7 % on the average), while the divergence among monomers within each HOR copy is sizeable (20 % on the average).[12] The studied region of the DNA molecule is fully noncoding. Here we present a detailed investigation of spectral densities for that central HOR domain.

The sequence of N nucleotides:

$$[n_i],\ i = 1,2,...N$$

(n_i denotes a nucleotide at the i-th position in the sequence) was transformed into the numerical sequence:

$$[u_i],\ i = 1,2,...N$$

using quartic mapping, by assigning different numbers to each of the four nucleotides:

$$\begin{aligned} u_i &= 4 \text{ if } x_i = \text{A} \\ u_i &= 3 \text{ if } x_i = \text{T} \\ u_i &= 2 \text{ if } x_i = \text{C} \\ u_i &= 1 \text{ if } x_i = \text{G}. \end{aligned}$$

In the second step, the fast Fourier transform (FFT) of each sequence $[u_i]$ was performed using FFT subroutine CO6EAF from the NAG library.[42] This computer routine calculates power spectra for discrete Fourier transforms using the $1/\sqrt{N}$ normalization. The applied hardware was PentiumIII and calculations were performed with double precision. The number of data included in the computation was taken in the standard way[42] as the value of the product of prime numbers (none exceeding 19) that is closest to the length of the sequence. To calculate the Fourier transform of the genomic sequence in the central domain of AC017075.8, we used $N = 2 \times 2 \times 3 \times 5 \times 11 \times 13 \times 17 = 145860$.

Tables I–III display the results for spectral densities of low-frequency, high-frequency and multiple monomer-frequency peaks in the Fourier spectrum of the central segment of the genomic sequence AC017075.8.

DISCUSSION

The Fourier spectrum up to frequency $f = 0.017558$ bp^{-1} exhibits peaks presented in Table I. The frequencies are denoted $f_1, f_2, f_3,...$ in the order of peak appearance. It is apparent that the frequencies of all peaks roughly correspond to a multiple pattern.

The lowest peak in the Fourier spectrum lies at frequency $f_1 = 0.000363$ bp^{-1} with spectral density $S_f = 2.7$. This frequency is slightly higher than the frequency corresponding to the 2734-bp HOR length

$$f_{\text{HOR}} = \frac{1}{2734\,\text{bp}} = 0.000366\ \text{bp}^{-1}$$

Thus, the lowest peak frequency

$$f_1 = 0.000363\ \text{bp}^{-1}$$

approximately corresponds to the HOR frequency. The slight difference between them reflects the computational effect of the finite range of the HOR sequence (54 HORs). Due to the data set truncation and the associated precision limitation (7.6×10^{-6}), the sequence length corresponding to frequency f_1 is $l_1 = 1/f_1 = 2.7 \times 10^3$. A more precise value was determined employing higher multiples, as it will be shown shortly.

Inspection of the computed frequencies shows that all prominent peaks above the noise background in the power spectrum lie at approximate multiples of the lowest frequency f_1. We note that such an extremely regular pattern can be rarely found even in most regular dynamical systems appearing in physics and engineering.

Accordingly, we describe these Fourier frequencies f_n using the multiple frequency formula:

$$f_n = n \times f_1^{(0)},\ (n = 1, 2, 3,...) \tag{1}$$

The value of $f_1^{(0)}$ in this formula is determined from the peak with the highest spectral density in the Fourier spectrum of the genomic sequence AC017075.8:

$$f_1^{(0)} = \frac{f_{224}}{224} \tag{2}$$

where f_{224} is the Fourier frequency of the 224th peak, which has the highest spectral density. Fourier transform computation gives $f_{224} = 0.081935$ bp^{-1}, and therefrom:

$$f_1^{(0)} = 0.00036578 \text{ bp}^{-1} \tag{3}$$

This value, deduced from the most pronounced peak in the Fourier spectrum, reproduces the precise HOR length value:

$$\frac{1}{f_1^{(0)}} = 2734,\ \textit{i.e.},$$
$$f_1^{(0)} = f_{\text{HOR}} \tag{4}$$

This clearly shows that the exact HOR frequency of 1/2734 bp^{-1} plays the role of fundamental frequency for the whole Fourier spectrum. It also provides a hint about the basic role of frequency f_{224}, which corresponds to the sequence length:

$$\frac{1}{f_{224}} \approx 12$$

hinting at the prominent role of a 12-bp sequence in relation to the HOR structure. One might argue about its possible connection to DNA folding.

The multiple formula (1, 2) provides a very good description of the computed frequencies corresponding to all peaks in the Fourier spectrum (deviations are less than 1 %) (Tables I–III). Table I displays 48 lowest peaks in Fourier spectrum of AC017075.8. We have identified one thousand peaks in accordance with formula (1, 2).

As seen from Table I, some of more prominent low-lying peaks are at frequencies $f_2 = 2\,f_1^{(0)}$, $f_6 = 6\,f_1^{(0)}$, and $f_{16} = 16\,f_1^{(0)}$. The lengths of the corresponding genomic sequences are approximately 1367 bp, 456 bp, and 171 bp, respectively. The corresponding spectral densities S_f are 2.85, 5.63, and 61.66, respectively. Frequency f_{16} corresponds to the approximately 171-bp alpha satellite monomer. These peaks correspond to multiples of a frequency associated with the approximately 171-bp monomer. More precisely, the Fourier frequency $f_{16} = 0.005855$ corresponds to the monomer length of $1/f_{16} = 170.8$ bp. This is consistent with our previous finding that the HOR structure of AC017075.8 comprises ten 171-bp, four 170-bp, and two 172-bp.[12]

The HOR structure of the genomic sequence is clearly reflected in the spectral densities: starting from the lowest peak, each 16th peak has a local maximum of spectral density.

Peaks corresponding to the monomer ($n = 16$) and to its multiples ($n = 32 = 2 \times 16$, $n = 48 = 3 \times 16$) (Table I) exhibit pronounced local maxima of spectral density. In addition, we have introduced a relative strength, defined as the ratio of peak to noise spectral density. Here, the level of noise was determined in the neighborhood of the corresponding peak (last column in Table I). Relative strengths also show local maxima at the positions of monomer multiples.

Table II displays a segment of the high-frequency region of peaks in the Fourier spectrum, from the 896th to the 932th peak. Even in this high-frequency region, the multiple formula (1, 2) provides a very good approximation of Fourier frequencies. However, we find substantial deviations in the pattern of spectral densities: the maxima of spectral densities are split among several peaks, and mostly shifted from peaks $n = 16k$ towards $n = 16k + 1$ and $n = 16k + 2$. For example, the $n = 896$ peak (*i.e.,* $n = 16 \times 56$), has the spectral density $S_f = 16.006$, while the spectral densities for the next two peaks are higher, $S_f\,(897) = 66.643$ and $S_f\,(898) = 19.834$.

In the higher-frequency region, above $14f_{16}$, the spectral densities gradually decrease. In the Fourier spectra of coding DNA sequences for primates, two major peaks were previously found in the high-frequency region, corresponding to frequencies $f = 1/3$ bp^{-1} and $f = 1/9$ bp^{-1}, related to the codon structure.[16] In the present case of alpha satellite DNA, the peak at $f = 1/3$ bp^{-1} was not identified.[16] This result is entirely as predicted for noncoding sequences, since the peak at $f = 1/3$ bp^{-1} is caused by the codon structure, which is absent here.

Table III displays the $n = 16k$ peaks, *i.e.,* a subset of peaks corresponding to multiples of frequency corresponding to the approximately 171-bp monomer. These peaks are referred to as monomer-multiples. It is seen that the lowest 16 monomer-multiples are characterized by highest local spectral densities. On the other hand, for

TABLE I. Frequencies and spectral densities for all peaks identified in the power spectrum of the central segment (31209 to 179354) of the complete genomic sequence AC017075.8 up to frequency 0.017558 bp^{-1}(a)

Peak(b)	Frequency f_n (c)	$n \times f_0$ (d)	Length l_n (e)	Spectral density		Relative
n	bp^{-1}	bp^{-1}	bp	Peak(f)	Noise(g)	strength(h)
1	**0.000363**	**0.000366**	**2.7×10^3**	**2.708**	**0.0214**	**127**
2	0.000727	0.000732	1.4×10^3	2.853	0.0496	57
3	0.001104	0.001097	906	1.397	0.0458	31
4	0.001467	0.001463	681	4.383	0.0778	56
5	0.001831	0.001829	546	0.521	0.0437	12
6	0.002201	0.002195	454	5.632	0.1335	42
7	0.002544	0.002560	393	1.657	0.0372	45
8	0.002941	0.002926	340	3.003	0.0816	37
9	0.003284	0.003292	305	7.239	0.0781	93
10	0.003647	0.003658	274	4.348	0.0647	67
11	0.004011	0.004024	249	9.877	0.1087	91
12	0.004388	0.004389	228	0.838	0.0428	20
13	0.004751	0.004755	210	4.775	0.0696	69
14	0.005128	0.005121	195	8.287	0.0599	138
15	0.005492	0.005487	182	0.709	0.0300	24
16 = 1 × 16	**0.005855**	**0.005853**	**171**	**61.659**	**0.1521**	**405**
17	0.006218	0.006218	161	0.745	0.0263	28
18	0.006575	0.006584	152	2.957	0.0592	50
19	0.006959	0.006950	144	4.982	0.0903	55
20	0.007322	0.007316	137	5.781	0.1508	38
21	0.007692	0.007681	130	0.684	0.0237	29
22	0.008056	0.008047	124	17.769	0.1632	109
23	0.008419	0.008413	119	3.714	0.0352	106
24	0.008762	0.008779	114	0.504	0.0198	25
25	0.009139	0.009145	109	11.922	0.0589	202
26	0.009502	0.009510	105	22.261	0.2267	98
27	0.009872	0.009876	101	0.580	0.0421	14
28	0.010236	0.010242	98	11.682	0.2234	52
29	0.010599	0.010608	94	1.731	0.0543	32
30	0.010963	0.010973	91	5.435	0.1085	50
31	0.011340	0.011339	88	17.029	0.0550	310
32 = 2 × 16	**0.011703**	**0.011705**	**85**	**24.217**	**0.0716**	**338**
33	0.012066	0.012071	83	13.189	0.0413	319
34	0.012430	0.012437	80	2.395	0.0532	45
35	0.012807	0.012802	78	1.503	0.0390	39
36	0.013150	0.013168	76	0.654	0.0420	16
37	0.013513	0.013534	74	0.181	0.0219	8
38	0.013911	0.013900	72	18.056	0.1755	103
39	0.014274	0.014265	70	3.982	0.0783	51
40	0.014651	0.014631	68	4.728	0.0850	56
41	0.015014	0.014997	67	4.550	0.1094	42
42	0.015350	0.015363	65	14.564	0.2116	69
43	0.015714	0.015729	64	4.996	0.1409	35
44	0.016091	0.016094	62	2.730	0.0665	41
45	0.016454	0.016460	61	13.177	0.1386	95
46	0.016831	0.016826	59	0.336	0.0211	16
47	0.017195	0.017192	58	12.151	0.0455	267
48 = 3 × 16	**0.017558**	**0.017558**	**57**	**185.060**	**0.0440**	**4204**

(a) Bold: peaks with n being multiples of 16.

(b) Ordering number n of a peak in the Fourier spectrum in the order of appearance.

(c) Frequency f_n corresponding to the nth peak in the Fourier spectrum.

(d) Frequency f_n predicted by approximate Eqs. (1, 2).

(e) Length of the sequence corresponding to frequency f_n, $l_n = 1/f_n$.

(f) Spectral density corresponding to the peak at frequency f_n.

(g) Level of noise in the neighborhood of the peak at frequency f_n.

(h) Ratio of the maximum peak spectral density and noise spectral density at frequency f_n.

TABLE II. Frequencies and spectral densities at all identified peaks in the power spectrum of the central segment (31209 to 179354) of the complete genomic sequence AC017075.8 in the frequency interval 0.327746 bp^{-1} to 0.340909 bp^{-1} [(a)]

Peak No.	Frequency f_n	$n \times f_1^{(0)}$	Length l_n	Spectral density		Relative strength
n	bp^{-1}	bp^{-1}	bp	Peak	Noise	
896 = 56 × 16	**0.327746**	**0.327740**	**3.05**	**16.006**	**0.4103**	**39**
897	0.328102	0.328106	3.05	66.643	1.8933	35
898	0.328466	0.328472	3.04	19.834	0.6219	32
899	0.328843	0.328837	3.04	3.617	0.1904	19
900	0.329213	0.329203	3.04	0.781	0.0991	8
901	0.329563	0.329569	3.03	2.402	0.1082	22
902	0.329947	0.329935	3.03	1.791	0.0737	24
903	0.330310	0.330300	3.03	7.088	0.1685	42
904	0.330666	0.330666	3.02	1.641	0.1161	14
905	0.331030	0.331032	3.02	13.868	0.2892	48
906	0.331393	0.331398	3.02	11.526	0.2625	44
907	0.331756	0.331764	3.01	36.999	1.0832	34
908	0.332099	0.332129	3.01	7.420	0.5027	15
909	0.332511	0.332495	3.01	19.895	0.9107	22
910	0.332511	0.332861	3.00	26.151	0.8658	30
911	0.333230	0.333227	3.00	10.366	0.3907	27
912 = 57 × 16	**0.333594**	**0.333593**	**3.00**	**18.899**	**0.2584**	**73**
913	0.333957	0.333958	2.99	42.448	0.8858	48
914	0.334321	0.334324	2.99	56.528	0.9559	59
915	0.334698	0.334690	2.99	1.153	0.1525	8
916	0.335061	0.335056	2.98	2.047	0.2162	9
917	0.335438	0.335421	2.98	45.351	1.4242	32
918	0.335801	0.335787	2.98	1.392	0.1659	8
919	0.336137	0.336153	2.97	7.798	0.1828	43
920	0.336521	0.336519	2.97	0.859	0.0778	11
921	0.336898	0.336885	2.97	0.811	0.1260	6
922	0.337241	0.337250	2.97	43.254	1.0912	40
923	0.337605	0.337616	2.96	17.363	0.3931	44
924	0.337982	0.337982	2.96	5.110	0.2556	20
925	0.338366	0.338348	2.96	6.220	0.5514	11
926	0.338722	0.338713	2.95	35.799	1.0971	33
927	0.339085	0.339079	2.95	29.646	0.3997	74
928 = 58 × 16	**0.339456**	**0.339445**	**2.95**	**0.508**	**0.1701**	**3**
929	0.339812	0.339811	2.94	33.535	0.8917	38
930	0.340176	0.340177	2.94	66.092	2.7545	24
931	0.340546	0.340542	2.94	18.592	0.4890	38
932	0.340909	0.340908	2.93	3.536	0.2893	12

[(a)] For description see Table I.

the 17th monomer-multiple and for most of the monomer-multiples above the 26th one, the spectral density is shifted away from these peaks, mostly to the neighboring peaks at higher frequencies $n = 16k + 1$ and $n = 16k + 2$, similarly as illustrated for peaks in Table II, and with more pronounced fluctuations.

In the higher frequency region, above the monomer frequency f_{16}, within the set of multiple frequencies nf_1, we observe a prominent subset of frequencies that are multiples of the monomer frequency f_{16}

$$f_k = k \cdot f_{16},\ (k = 1, 2, 3,...)$$

The peaks corresponding to multiples of f_{16} are sizably stronger than the peaks corresponding to nf_1 for $n \neq 16k$ (k =1, 2, 3,...). Particularly pronounced frequencies in the power spectrum are $14\,f_{16}$, $12\,f_{16}$, $10\,f_{16}$ and $5\,f_{16}$, with the corresponding spectral densities 1668.7, 727.2, 650.5 and 555.1, respectively. The corresponding lengths of subsequences are approximately 12 bp, 14 bp, 17 bp, and 34 bp, respectively. This reveals a complex substructure of monomer repeats, *i.e.,* approximately conserved embedded subrepeats. The first pronounced higher harmonic of the monomer frequency 1/171 bp^{-1} is at approximately 1/57 bp^{-1}. Accordingly, we can sub-

TABLE III. Frequencies and spectral densities for monomer-multiple peaks (n = 16k, k = 1, 2, 3,...) in the power spectrum of the central segment (31209 to 179354) of the complete genomic sequence AC017075.8 for frequencies up to 0.362855 bp^{-1} [(a)(b)]

$k = \left(\frac{n}{16}\right)$	Frequency f_n	$k \times 16f_1^{(0)}$	Length l_n	Spectral density		Relative strength
	bp^{-1}	bp^{-1}	bp	Peak	Noise	
1	**0.005855**	**0.005853**	**171**	**61.659**	**0.1521**	**405**
2	**0.011703**	**0.011705**	**85**	**24.217**	**0.0716**	**338**
3	**0.017558**	**0.017558**	**57**	**185.060**	**0.0440**	**4204**
4	**0.023413**	**0.023410**	**43**	**230.610**	**0.7646**	**301**
5	**0.029261**	**0.029263**	**34**	**555.150**	**0.2148**	**2585**
6	**0.035116**	**0.035115**	**28.5**	**271.361**	**0.2031**	**1336**
7	**0.040964**	**0.040968**	**24.4**	**31.507**	**0.0970**	**325**
8	**0.046819**	**0.046820**	**21.4**	**216.149**	**0.1047**	**2066**
9	**0.052674**	**0.052673**	**19.0**	**99.611**	**0.1647**	**605**
10	**0.058522**	**0.058525**	**17.1**	**650.538**	**0.6373**	**1020**
11	**0.064377**	**0.064378**	**15.5**	**80.581**	**0.0773**	**1042**
12	**0.070232**	**0.070230**	**14.2**	**727.228**	**1.5984**	**455**
13	**0.076080**	**0.076083**	**13.1**	**466.188**	**0.4569**	**1020**
14	**0.081935**	**0.081935**	**12.2**	**1668.700**	**2.2053**	**757**
15	**0.087783**	**0.087788**	**11.4**	**177.590**	**0.5071**	**350**
16	**0.093638**	**0.093640**	**10.7**	**80.722**	**0.1110**	**727**
17	0.099493	0.099493	10.1	16.500	0.0855	193
18	**0.105341**	**0.105345**	**9.5**	**282.260**	**0.7711**	**366**
19	**0.111196**	**0.111198**	**9.0**	**376.230**	**0.6857**	**549**
20	**0.117051**	**0.117050**	**8.5**	**179.090**	**0.9219**	**194**
21	**0.122899**	**0.122903**	**8.1**	**47.099**	**0.1779**	**265**
22	**0.128754**	**0.128755**	**7.8**	**74.392**	**0.3559**	**209**
23	**0.134602**	**0.134608**	**7.4**	**54.551**	**0.4143**	**132**
24	**0.140457**	**0.140460**	**7.1**	**29.752**	**0.1584**	**188**
25	**0.146312**	**0.146313**	**6.8**	**170.290**	**1.1730**	**145**
26	**0.152160**	**0.152165**	**6.6**	**178.320**	**1.2159**	**147**
27	0.158015	0.158018	6.3	1.043	0.0395	26
28	0.163876	0.163870	6.1	165.730	0.9796	169
29	**0.169718**	**0.169723**	**5.9**	**79.921**	**0.6380**	**125**
30	**0.175572**	**0.175575**	**5.7**	**79.863**	**0.6682**	**120**
31	**0.181434**	**0.181428**	**5.5**	**62.232**	**0.3634**	**171**
32	0.187275	0.187280	5.3	20.926	0.2303	91
33	0.193137	30.193133	5.2	19.957	0.2027	98
34	0.198992	0.198985	5.0	94.943	0.8541	111
35	0.204833	0.204838	4.9	72.251	0.9614	75
36	0.210695	0.210690	4.7	1.082	0.0619	17
37	0.216550	0.216543	4.6	39.982	0.4547	88
38	**0.222398**	**0.222395**	**4.5**	**31.945**	**0.3737**	**85**
39	0.228253	0.228248	4.4	175.999	1.3244	133
40	0.234108	0.234100	4.3	4.183	0.1185	35
41	0.239956	0.239953	4.2	122.779	1.2707	97
42	0.245811	0.245805	4.1	3.089	0.0923	33
43	0.251659	0.251658	4.0	16.753	0.3393	49
44	0.257514	0.257510	3.9	54.624	0.5084	107
45	**0.263369**	**0.263363**	**3.8**	**25.609**	**0.2940**	**87**
46	0.269217	0.269215	3.7	15.487	0.3416	45
47	0.275072	0.275068	3.6	60.266	0.4983	121
48	0.280927	0.280920	3.56	65.248	1.3980	47
49	0.286775	0.286773	3.49	26.259	0.4584	57
50	0.292630	0.292625	3.42	0.973	0.1369	7
51	0.298485	0.298478	3.35	0.306	0.0535	6

TABLE III. cont.

$k = \left(\frac{n}{16}\right)$	Frequency f_n	$k \times 16f_1^{(0)}$	Length l_n	Spectral density		Relative strength
	bp^{-1}	bp^{-1}	bp	Peak	Noise	
52	0.304333	0.304330	3.29	2.077	0.0808	3
53	0.310188	0.310183	3.22	11.678	0.2002	58
54	0.316036	0.316035	3.16	14.706	0.3552	41
55	0.321891	0.321888	3.11	3.564	0.1614	22
56	0.327746	0.327740	3.05	16.006	0.4103	39
57	0.333594	0.333593	3.00	18.899	0.2584	73
58	0.339456	0.339445	2.95	0.508	0.1701	3
59	0.345311	0.345298	2.90	0.508	0.0628	8
60	0.351152	0.351150	2.85	17.740	0.3119	57
61	0.357007	0.357003	2.80	3.019	0.0842	36
62	0.362855	0.362855	2.75	3.450	0.1365	25

[a] Results are presented for peaks number $n = 16k$ (k = 1, 2, 3,...).
[b] Bold: peaks that have locally the highest spectral density (highest by comparison with several neighboring peaks $n \neq 16k$ in the Fourier spectrum).

divide the 171-bp monomer into three approximately 57-bp subrepeats. More precisely, our direct inspection of the genomic sequence has shown that the 171-bp monomer is subdivided into three variants of approximately 57-bp subrepeats,

$$171 \text{ bp} = 56 \text{ bp} + 57 \text{ bp} + 58 \text{ bp}.$$

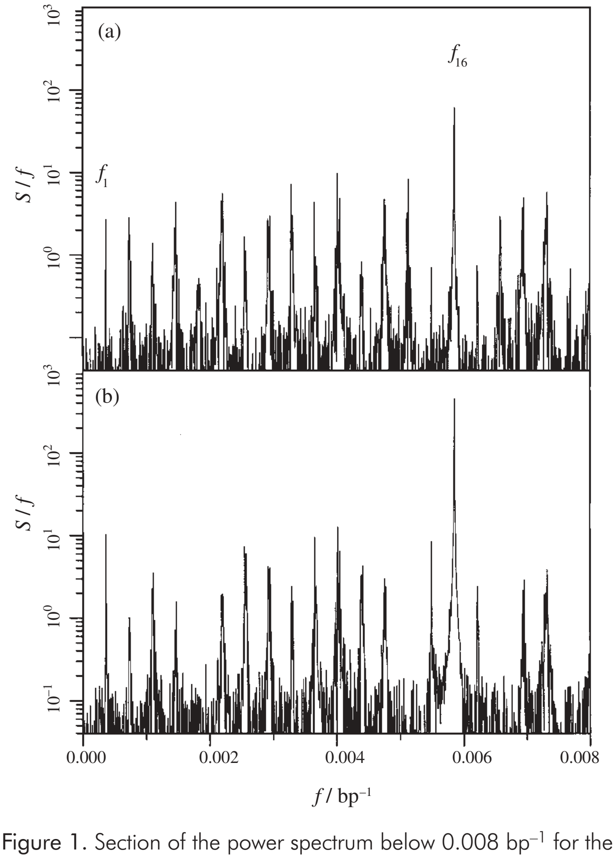

Figure 1. Section of the power spectrum below 0.008 bp^{-1} for the clone AC017075.8 in human chromosome 7. (a) Quartic mapping A → 4, T → 3, C → 2, G → 1. (b) Quartic mapping A → 4, T → 2, C → 3, G → 1.

Let us comment on the impact of our choice of numerical assignment in quartic mapping used in calculations. We employed a mapping with the nucleotide pairs A,T and G,C corresponding to pairs of the neighboring integers 4,3 and 2,1, respectively. Tables I–III present the resulting power spectra. Diagrammatic presentation of the segment below 0.008 bp^{-1} is additionally displayed in Figure 1(a). For comparison, in Figure 1(b) we display the power spectra with a different choice of numerical assignments, where A,T and C,G are assigned to integers 4,2 and 3,1, respectively. As seen from the comparison of Figures (a) and (b), the peak frequencies appear robust, only some strengths are modified.

Finally, we note that the power spectrum of the central segment of AC017075.8, with a very long sequence of equidistant peaks, shows an extremely pronounced pattern resembling frequency locking. Frequency locking is a well-known phenomenon that appears in natural sciences and engineering.[43] If there are two competing fundamental frequencies with a rational ratio, and if the interaction includes a term as a product of circular functions, then all peaks in the Fourier spectrum are harmonics (multiples) of a single frequency f_1, built as a specific linear combination of two fundamental frequencies.[43] In that case, the Fourier spectrum is equidistant, and the peak corresponding to the lowest frequency is usually not the strongest one.

CONCLUSIONS

We have found a characteristic multiple-frequency pattern for the higher-order repeat 16mer in human alpha satellite DNA in chromosome 7, with the HOR frequency having the role of fundamental frequency. Additionally, a hierarchy of periodicities in the monomer sequence was identified.

We can conclude that mutations, insertions and deletions imposed on the ideal HOR structure (consensus HOR) have only a minor impact on the multiple-frequency pattern, which resists these deviations, while the spectral density pattern in the high-frequency region of the Fourier spectrum is more sensitive, resulting in a shift of spectral density away from monomer-multiples frequencies and its splitting among several near-lying peaks.

An important conclusion that follows from the comparison of spectral densities is the dominant role exhibited by the 2734-bp HOR sequence, which can be deduced from the 14th monomer-multiple (n = 14 × 16 = 224).

Finally, we note that the Fourier transform provides a global method, rather insensitive to smaller deviations of periodicity, for identifying the HOR and internal monomer structure in a given genomic sequence. Once this is established for a particular sequence, an exact analysis determining in detail all mutations, deletions and insertions can be performed using the recently introduced Key-string algorithm.[12]

We propose similar structural investigations for the centromeric regions of all chromosomes as well as determination of the corresponding fundamental frequencies.

REFERENCES

1. L. Manuelidis, *Chromosoma* **66** (1978) 23–32.
2. J. S. Waye and H. F. Willard, *Nucleic Acids Res.* **15** (1987) 7549–7569.
3. H. F. Willard and J. S. Waye, *J. Mol. Evol.* **25** (1987) 207–214.
4. J. S. Waye and H. F. Willard, *Chromosoma* **98** (1989) 273–279.
5. H. F. Willard, *Trends Genet.* **6** (1990) 410–416.
6. R. Wevrick, V. P. Willard, and H. F. Willard, *Genomics* **14** (1992) 912–923.
7. A. de la Puente, E. Velasco, L. A. Perez Jurado, C. Hernandez Chico, F. M. van de Rijke, S. W. Scherer, A. K. Raap, and J. Cruces, *Cytogenet. Cell. Genet.* **83** (1998) 176–181.
8. P. Vogt, *Hum. Genet.* **84** (1990) 301–336.
9. C. Lee, R. Wevrick, R. B. Fisher, M. A. Ferguson-Smith, and C. C. Lin, *Hum. Genet.* **100** (1997) 291–304.
10. J. S. Waye, S. B. England, H. F. Willard, and H. F. Willard, *Mol. Cell Biol.* **7** (1987) 349–356.
11. R. Wevrick and H. F. Willard, *Nucleic Acids Res.* **19** (1991) 2295–2301.
12. M. Rosandić, V. Paar, and I. Basar, *J. Theor. Biol.* **221** (2003) 29–36.
13. R. Waterston, GenBank accession no. AC017075.8 (2002).
14. N. Nagai, K. Kuwata, T. Hayashi, H. Kuwata, and S. Era, *Jpn. J. Physiol.* **51** (2001) 159–168.
15. W. Li and K. Kaneko, *Europhys. Lett.* **17** (1992) 655–660.
16. R. F. Voss, *Phys. Rev. Lett.* **68** (1992) 3805–3808.
17. B. Borstnik, D. Pumpernik, and D. Lukman, *Europhys. Lett.* **23** (1993) 389–394.
18. V. R. Chechetkin and A. Y. Turygin, *J. Theor. Biol.* **175** (1995) 477–494.
19. E. Coward, *J. Math. Biol.* **36** (1997) 64–70.
20. S. Tiwari, S. Ramachandran, S. Bhattacharya, and R. Ramaswami, *Comp. Appl. Biosci.* **13** (1997) 263–270.
21. G. I. Kutuzova, G. K. Frank, V. Y. Makeev, N. G. Esipova, and R. V. Polozov, *Biofizika* **42** (1999) 354–362.
22. C. M. Pasquier, V. I. Promponas, N. J. Varvayannis, and S. J. Hamodrakas, *Bioinformatics*, **14** (1998) 749–750.
23. M. Osaka, K. Gohara, S. Ishii, H. Kishida, H. Hayakawa, and N. Ito, *Physica D* **125** (1999) 142–154.
24. S. Guharay, B. R. Hunt, J. A. Yorke, and O. R. Whitew, *Physica D* **146** (2000) 388–396.
25. Z.-G. Yu, V. Anh, and K.-S. Lau *Phys. Rev. E* **64** (2001) 031903, 1–9.
26. E. Pizzi, S. Liuni, and C. Frontali, *Nucleic Acids Res.* **18** (1990) 3745–3752.
27. C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. Simons, and H. E. Stanley, *Physica A* **221** (1995) 180–192.
28. S. Nee, *Nature* **357** (1992) 450.
29. C. A. Chatzidimitriou-Dreismann and D. Larhammar, *Nature* **361** (1993) 212–213.
30. C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, *Nature* **356** (1992) 168–170.
31. J. S. Almeida, J. A. Carrico, A. Maretzek, P. A. Noble, and M. Fletcher, *Bioinformatics* **17** (2001) 429–437.
32. A. Arneodo, Y. d'Aubenton-Carafa, B. Audit, E. Bacry, J. F. Muzy, and C. Thermes, *Eur. Phys. J. B* **1** (1998) 259–263.
33. M. F. Sagot and E. W. Myers, *J. Comput. Biol.* **5** (1998) 539–553.
34. J. H. Jackson, R. George, and P. A. Herring, *Biochem. Biophys. Res. Commun.* **268** (2000) 289–292.
35. Z.-G. Yu and P. Jiang, *Phys. Lett. A* **286** (2001) 34–46.
36. V. V. Lobzin and V. R. Chechetkin, *Uspekhi Fiz. Nauk* **170** (2000) 57–81.
37. P. Bernaola-Galvan, R. Roman–Roldan, and J. L. Oliver, *Phys. Rev. E* **53** (1996) 5181–5189.
38. B. Borstnik, D. Pumpernik, D. Lukman, Đ. Ugarković, and M. Plohl, *Nucleic Acids Res.* **22** (1994) 3412–3417.
39. B. Borstnik and D. Pumpernik, *Genome Res.* **12** (2002) 909–915.
40. V. R. Chechetkin, L. A. Knizhnikova, and A. Y. Turygin, *J. Biomol. Struct. Dyn.* **12** (1994) 271–299.
41. N. G. Esipova, G. I. Kutuzova, V. Y. Makeev, G. K. Frank, A. V. Balandina, D. E. Kamashev, and V. L. Karpov, *Biofizika* **45** (2000) 432–438.
42. *NAG Fortran Library*, Oxford NAG Ltd., Oxford (1990).
43. P. Berge, Y. Pomeau, and C. Vidal, *Order Within Chaos*, Wiley, New York, 1984.

SAŽETAK

Spektralne gustoće i frekvencije u Fourierovom spektru repeticija alfa satelita višega reda u humanoj DNA molekuli

Vladimir Paar, Nenad Pavin, Ivan Basar, Marija Rosandić, Ivica Luketin i Sonja Durajlija Žinić

Brza Fourierova transformacija primijenjena je na središnji dio potpuno sekvenciranoga genomskoga segmenta iz područja centromere u humanom kromozomu 7 (GenBank / AC017075.8, 193277 bp), koji je karakteriziran alfa satelitskom repeticijom višega reda (HOR). Frekvencije i spektralne gustoće su izračunane za sve istaknute maksimume u Fourierovom spektru. Dodatno je uveden kvocijent spektralne gustoće maksimuma i šuma kao efektivna spektralna gustoća, kako bi se u obzir uzela varijacija frekvencije razine šuma. Pokazano je da se izvrstan opis izračunanih Fourierovih frekvencija dobije pomoću multipolne formule u kojoj fundamentalna frekvencija odgovara HOR-u s 2734 baza. Maksimum za frekvenciju f_{16} odgovara monomeru sa 171 bazom. Iznad frekvencije f_{16} najizraženiji maksimumi su višekratnici f_{16} (monomer-multipleti). Šesnaest najnižih monomer-multipleta kf_{16} lokalno su dominantni u spektralnim gustoćama. Najniži monomer-multiplet koji nije lokalno dominantan u spektralnoj gustoći javlja se za k = 17. Iznad k = 27 maksimumi spektralne gustoće sustavno su pomaknuti prema nekoliko susjednih viših frekvencija. Na temelju Fourierovoga spektra, struktura monomerne jedinice sa 171 bazom fragmentirana je u tri aproksimativno 57-baznih pod-repeticija koje se zatim fragmentiraju u 12-bazne, 14-bazne i 17-bazne pod-repeticije.