CABIOS

Vol. 13 no. 1 1997
Pages 37–44

Latent sequence periodicity of some oncogenes and DNA-binding protein genes

E.V.Korotkov, M.A.Korotkova[1] and J.S.Tulko

Abstract

A method of latent periodicity search is developed. We use mutual information to reveal the latent periodicity of mRNA sequences. The latent periodicity of an mRNA sequence is a periodicity with a low level of similarity between any two periods inside the mRNA sequence. The mutual information between an artificial numerical sequence and an mRNA sequence is calculated. The length of the artificial sequence period is varied from 2 to 150. The high level of the mutual information between artificial and mRNA sequences allows us to find any type of latent periodicity of mRNA sequence. The latent periodicity of many mRNA coding regions has been found. For example, the retinoblastoma gene of HSRBS clone contains a region with a latent period equal to 45 bases. The A-RAF oncogene of HSARAF1R clone contains a region with a latent period equal to 84 bases. Integrated sequences for the regions with latent periodicity are determined. The potential significance of latent periodicity is discussed.

Introduction

The search for periodical sequences in DNA and RNA is important for understanding the DNA sequence structure of genomes of the human and other species. Periodical sequences in the human genome are now found as satellite and minisatellite sequences (Bliskovsky, 1991; Hearne *et al.*, 1992; Todd, 1992). These sequences contain a short repeated fragment of DNA.

Mathematical methods of determining symbolical sequence periodicity have now been developed (McLachlan, 1977b, 1993; Cheever *et al.*, 1991; Voss, 1992; Chechetkin *et al.*, 1994; McLachlan and Stewart, 1994). These approaches are directed to the application of mathematical methods advanced for analysis of periodicity in numerical sequences to study the periodicity of symbolical sequences. A transformation of a symbolic sequence to a numerical sequence is frequently used for such analysis. The investigations have shown a periodicity of amino acid sequences of some genes (Stewart and McLachlan, 1975; McLachlan, 1977a,b, 1993; McLachlan and Stewart, 1994) and a periodicity of some DNA sequences (Pizzi *et al.*, 1990; Cheever *et al.*, 1991; Bina, 1994; Borstnik *et al.*, 1994; Chechetkin *et al.*, 1994).

The study of the periodicity of DNA sequences is mainly based on the analysis of homological periodicity including imperfect repeats (Silverman and Linsker, 1986; von Heijne, 1987; Voss, 1992; Makeev and Tumanyan, 1994). However, hidden periodicity of a DNA sequence may exist and this periodicity cannot be found by the algorithms developed. For example, there may be observed a pattern (A/C/T)(T/G)(G/A)(T/A)(C/G/A)(G/T) where the first position of a period contains A, C or T, the second position contains T or G, and so on. Earlier, a new mathematical method of detecting latent periodicity was developed and the latent periodicity of some human genes was found (Korotkov and Korotkova, 1995).

The mathematical approach directed at detecting similar features in DNA and RNA sequences is improved in the present work. The latent periodicity of some mRNA regions is shown. These mRNA contain the following genes: a retinoblastoma gene, a A-RAF oncogene, GLI-protein, transcription factor SP1, human homeobox C1 and C8 genes, an ABL-oncogene, an FMS-oncogene, an epidermal growth factor receptor gene. The mathematical approach being developed uses the method of enlarged similarity of DNA sequences (Korotkov and Korotkova, 1993; Korotkov, 1994).

System and methods

A comparison of artificial periodic sequences with the RNA sequence analyzed is used for the detection of latent periodicity (Korotkov and Korotkova, 1995). The alphabet of artificial sequences contains letters S_i. The sequence $S_1S_2S_1S_2S_1S_2$... is generated for the search for an mRNA period equal to two bases. The sequence $S_1S_2...S_nS_1S_2...S_nS_1S_2...S_n$... is generated for the study of the mRNA period equal to n bases. The length of the artificial sequence is equal to the length of the analyzed sequence. The artificial sequences having a period from two up to 150 are serially compared with the analyzed mRNA sequence. Mutual information is chosen as a measure of the similarity of artificial and mRNA sequences. A matrix M is filled in for calculation of the mutual information. Elements of the matrix M are the numbers

Center of Bioengineering of the Russian Academy of Sciences, 60-Oktybrya prospect, 7/1, Moscow, 117312 and Moscow Physical Engineering Institute, Department of Cybernetics, Ministry of Education of Russia, 115409, Moscow, Russia

[1]*To whom correspondence should be addressed at: Moscow Physical Engineering Institute, Department of Cybernetics, Ministry of Education of Russia, Kashirskoe shosse, 31, 115409, Moscow, Russia*

of coincidences of each type between the artificial sequence and the mRNA sequence. The dimension of matrix M is $4 \times n$. The labels of the matrix M rows are the bases A, U, C and G. Labels of the columns of the matrix M are the letters S_i ($i = 1, 2, ..., n$) of the artificial sequence. The sums of row elements are equal to the quantity of A, U, C and G bases in the mRNA sequence. The sums of elements in each column are equal to the quantity of S_i letters in the artificial sequence. The mutual information is calculated using the formula (Kullback, 1959):

$$I = \sum_{1}^{4}\sum_{1}^{n} \mathrm{m}_{ij} \ln \mathrm{m}_{ij} - \sum_{1}^{4} \mathrm{x}_i \ln \mathrm{x}_i - \sum_{1}^{n} y_i \ln \mathrm{y}_j - L \ln L \quad (1)$$

Here, m_{ij} is the element of a matrix M, x_i is the quantity of each A, U, C and G symbol in the mRNA sequence, y_j is the quantity of the S_j symbol in the artificial sequence and L is the length of compared sequences. The $2I$ value is distributed as χ^2 with $3(n - 1)$ degrees of freedom. It permits us to evaluate the probability of accidental formation of periodicity.

We tested the conformity of the $2I$ distribution to the χ^2 distribution with $3(n - 1)$ degrees of freedom. We compared artificial sequences of different lengths with 5×10^6 base pairs of random sequence. The random sequence was generated from language text by the amalgamation of letters into four letters and mixing of letters. If the number of periods in the artificial sequence (each period had n bases) was more than four, then the $2I$ distribution corresponds to the χ^2 distribution with $3(n - 1)$ degrees of freedom with probability no less than 85%. This probability was >90% if the number of periods in the artificial sequence was greater than or equal to six.

All possible periods can be divided into two classes. The first class (prime periods) includes periods that are equal to prime numbers. The second class (complex periods) contains periods equal to a product of prime numbers. Let us have a compound period $\mathrm{A} = \mathrm{B} \times \mathrm{C}$, where B and C are prime periods. Let the period A have the least probability $\alpha = P(\chi^2 \geq 2I)$ for all periods analyzed. It was shown earlier (Yaglom and Yaglom, 1960):

$$I(\mathrm{RNA}, \mathrm{B}) + I_{\mathrm{RNA}}(\mathrm{C}, \mathrm{B}) + I(\mathrm{RNA}, \mathrm{C}) = I(\mathrm{C} \times \mathrm{B}, \mathrm{RNA}) + I(\mathrm{B}, \mathrm{C}) \quad (2)$$

Here, I(RNA,B) is the mutual information between the RNA sequence and the artificial sequence with the period that is equal to B, I(RNA,C) is the mutual information between the RNA sequence and the artificial sequence with the period that is equal to C, I(B,C) is the mutual information between two artificial sequences with periods that are equal to B and C and I(B × C,RNA) is the mutual information between the RNA sequence and a compound artificial sequence with the period that is equal to B × C. For example, let us have B = 3 and C = 2. We may write the sequences B and C as:

$\mathrm{b_1b_2b_3b_1b_2b_3b_1b_2b_3b_1b_2b_3}$...
$\mathrm{c_1c_2c_1c_2c_1c_2c_1c_2c_1c_2c_1c_2}$...

Then we may unite those two sequences and introduce six new letters: $\mathrm{a_1 = b_1c_1}$; $\mathrm{a_2 = b_2c_2}$; $\mathrm{a_3 = b_3c_1}$; $\mathrm{a_4 = b_1c_2}$; $\mathrm{a_5 = b_2c_1}$; $\mathrm{a_6 = b_3c_2}$. The sequence $\mathrm{A} = \{\mathrm{a_1a_2a_3a_4a_5a_6a_1a_2\text{-}a_3a_4a_5a_6a_1a_2a_3a_4a_5a_6}...\}$ is a united sequence B × C. The same method is applied if we construct the sequences (RNA × B), (RNA × C) and (RNA × C × B).

The mutual information I(B,C) is calculated as:

$$I(\mathrm{B}, \mathrm{C}) = \{H(\mathrm{B}) + H(\mathrm{C}) - H(\mathrm{B} \times \mathrm{C})\}L \quad (3)$$

where H(B), H(C) and H(B × C) are the entropies of B, C and B × C sequences, and L is the length of sequences. The H(B) is calculated as:

$$H(\mathrm{B}) = \Sigma p_i(\mathrm{B})\{\ln p_i(\mathrm{B})\} \quad (4)$$

where p_i(B) is the probability of letter i in B sequence ($i = 1, 2, ...$ B). The entropies H(C) and H(B × C) are calculated in a similar way. I_{RNA}(C,B) is the conditional mutual information between C and B sequences (Yaglom and Yaglom, 1960).

$$I_{\mathrm{RNA}}(\mathrm{C}, \mathrm{B}) = H(\mathrm{RNA} \times \mathrm{B}) + H(\mathrm{RNA} \times \mathrm{C}) - H(\mathrm{RNA}) - H(\mathrm{RNA} \times \mathrm{C} \times \mathrm{B}) \quad (5)$$

Here, H(RNA × B) is the entropy of the united sequence RNA × B, H(RNA × C) is the entropy of the united sequence RNA × C, H(RNA) is the entropy of the DNA sequence and H(RNA × C × B) is the entropy of the united RNA × C × B sequence. I_{RNA}(C,B) is not negative (Yaglom and Yaglom, 1960). If periods of the two sequences B and C are represented by prime numbers, I(B,C) = 0. It means that the mutual information I(RNA,B) and I(RNA,C) are independent parts of I(RNA,A). If the mutual information is considered as a random value, then I(RNA,B) and I(RNA,C) are independent random values. According to Kullback (1959), $2I$(RNA,B) and $2I$(RNA,C) are distributed as χ^2 with 3(B − 1) and 3(C − 1) degrees of freedom accordingly.

The mutual information I(RNA,kB), where $k = 2, 3, 4 ...$, may be equal to or exceed I(RNA,B). For example, if we calculate the mutual information between the artificial sequences with the periods of 3, 6, 9, 12 ... and RNA, then the mutual information for periods 6, 9, 12 ... is greater than or equal to I(RNA,3). This means that the mutual information is accumulated when k is increased.

For the compound period A, it is convenient to take the value I_{RNA}(C,B) as a measure of the sequence periodicity. This double value has the distribution χ^2 with number of degrees of freedom equal to 3(A − 1) − 3(B − 1) − 3(C − 1). I_{RNA}(C,B) reflects a contribution of the period A to the creation of this periodicity by excluding all simple influence. It is convenient to obtain the I_{RNA}(C,B) spectrum for any period in practice. It may be done by serial deduction of the mutual information of prime periods from the mutual information of compound periods. It is possible to carry out a similar subtraction for degrees of freedom, because the mutual information is distributed as χ^2 with $3(n-1)$ degrees of freedom (where n is the period length). Then mutual pieces of information of compound periods are deducted from those of longer divisible compound periods. For example, the mutual information $I(\mathrm{RNA},6)$ is deducted from $I(\mathrm{RNA},12)$, $I(\mathrm{RNA},18)$, and so on. The obtained spectrum of modified mutual information $I_m(n)$ shows the contribution of each period n to formation of periodicity. The values $2I_m(n)$ are distributed as χ^2 with the remaining number of degrees of freedom.

A contribution of each period in the observed periodicity can be conveniently obtained from the spectrum $I_m(n)$. For example, if an RNA sequence has periodicity only in 18 bases, then the modified $I_m(n)$ spectrum has only one maximum. If an RNA sequence has periods equal to 3, 6, 9 and 18 bases, then the $I_m(n)$ spectrum has four comparatively small maxima. The modified $I_m(n)$ spectrum may be applied for the detection of fractional periods.

The modified spectrum $I_m(n)$ must be represented as a spectrum of argument x of the normal distribution for evaluation of the statistical importance of each period. Using equation (6), transformation of $I_m(n)$ is performed (*Handbook of Applicable Mathematics*, Volume IV, 1984):

$$X(n) = \sqrt{4I_m(n)} - \sqrt{2t-1} \qquad (6)$$

Here, t is the number of degrees of freedom for $I_m(n)$. This spectrum $X(n)$ will be shown below for some RNA regions with latent periodicity.

As far as the RNA sequences of coding regions have period equal to three bases (von Heijne, 1987), the search for latent periodicity should be conducted by deducting $I(\mathrm{RNA},3)$ from all divisible periods. It allows elimination of the influence of triplet periodicity on long latent periodicity. The search for RNA regions with $I(\mathrm{RNA},3n) - I(\mathrm{RNA},3)$, which normal distribution argument exceed 5.6, is conducted in the present work. A value of 5.6 corresponds to a probability of 10^{-8} that the long periodicity found is caused by random factors.

Results and discussion

The search for regions with latent periodicity was performed in RNA clones from the EMBL data bank. Clones of length less than 1000 bases were not analyzed. An artificial sequence containing 1000 bases was compared with the first 1000 bases of each RNA clone. Independent variations of the left and right borders were conducted for each artificial sequence with a period from two up to 150. The purpose was to find an RNA region having the best periodicity and the maximum of $I(\mathrm{RNA},3n) - I(\mathrm{RNA},3)$.

If an RNA region with latent periodicity was not found, then displacement of the scanned artificial sequence by 100 bases was performed. If a region with latent periodicity was revealed, then the displacement of the artificial sequence by 500 bases was conducted, and the procedure of calculations was repeated. The full length of a clone was analyzed in such a way. After that, the next RNA clone from the EMBL data bank was analyzed.

The results of the analysis reveal many RNA regions with latent periodicity. More than 30% of human mRNAs (with length >1000 bases) from the EMBL data bank have regions with latent periodicity. Nine regions with the most strongly expressed latent periods are shown on Figure 1 and Table I. The spectra of $X(n)$ for regions represented on Figure 1 are shown on Figure 2. $2I'(\mathrm{RNA},n)$ values for these regions are adduced in Figure 1 and Table II. The $2I'(n,\mathrm{RNA})$ is equal to $2I(\mathrm{RNA},n) - 3(n-1)$. The significance $3(n-1)$ is an average significance of $2I(\mathrm{RNA},n)$ when an artificial sequence and an RNA sequence are the random sequences.

The sequences found with latent periodicity (except HSTFSP1 and HSGLI clones) have the significant period equal to three bases. These seven mRNA sequences have the $2I(3,\mathrm{RNA})$ in the interval of 20 up to 110. This corresponds to the probability α from $<10^{-9}$ up to 10^{-4}. Clones HSTFSP1

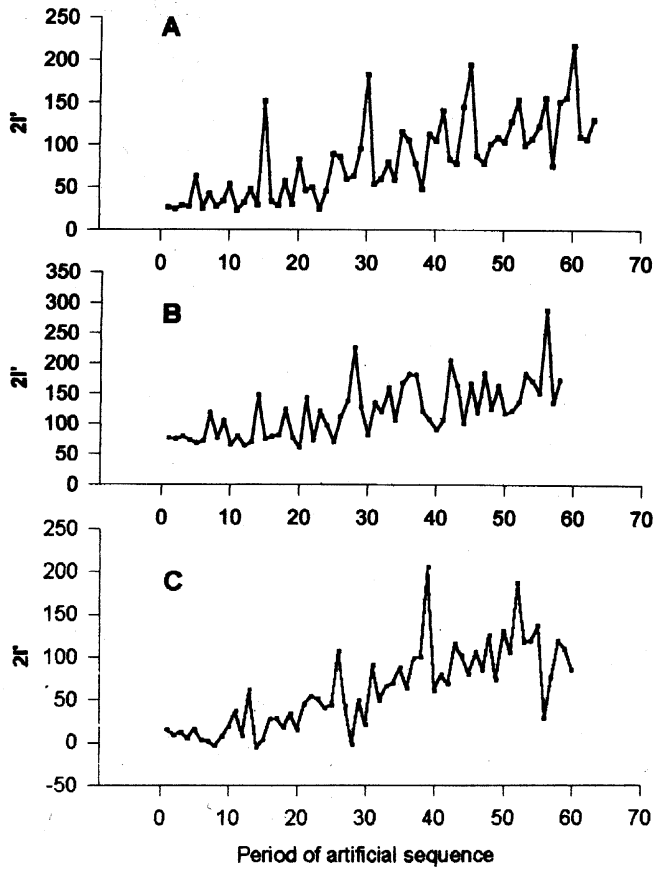


Fig. 1. The dependence of $2I'$ upon period of artificial sequence. The length of period is shown as divisible by three. (**A**) mRNA region (1022–1771 bases) from HSRBS clone (Lee *et al.*, 1987). (**B**) mRNA region (1101–1797 bases) from HSARAF1R clone (Beck *et al.*, 1987). (**C**) mRNA region (2725–3454 bases) from HSGLI clone (Kinzler *et al.*, 1988).

Table I. The coordinates of coding regions with latent periodicity in EMBL clones

Clones from EMBL data bank	Genes where the latent periodical sequences are found	Coordinates of found sequences in EMBL clones	Period length (bases)
HSTFSP1	transcription factor SP1	493–1096	105
HSHOXB	homeobox gene C8	282–973	111
HSHOX2I	homeobox gene C8	113–982	111
HSABL	P150 protein	858–1815	87
HSCFMS	proto-oncogene FMS	2051–2922	99
HSEGFRS	epidermal growth factor receptor	1338–2192	111

and HSGLI have $2I(3,\text{RNA})$ equal to 16 which corresponds to the probability α equal to 10^{-3}.

Against the background of the period equal to three bases (which is characteristic of DNA coding regions), periods of a size divisible by three are found. The values $2I'(3n,\text{RNA})$ for three regions from HSRBS (Lee *et al.*, 1987), HSARAF1R (Beck *et al.*, 1987) and HSGLI (Kinzler *et al.*, 1988) clones are shown in Figure 1. These regions are components of RB oncogene, A-RAF oncogene and GLI protein. The region with latent periodicity of RB oncogene has the period equal to 45 bases and contains half of the RB gene. HSARAF1R clone has the region with latent period that is equal to 84 bases. This region contains about a half of an A-RAF oncogene. GLI protein contains a zinc 'finger' region and may interact with DNA. The periodicity in 117 bases in the HSGLI clone is observed outside the zinc 'finger' region. This region with latent periodicity contains about one-third of the GLI protein.

The similarity between any pair of periods of the HSRBS clone is statistically insignificant. Such unimportant similarity is also observed between periods of HSARAF1R and HSGLI clones. The periods of the HSRBS clone are shown in Figure 3. Values $\chi^2(l)$ are shown above the sequences .

$$\chi^2(l) = \sum_{1}^{4}(n_i(l) - p_iN)^2/p_iN \qquad (7)$$

Here, $n_i(l)$ is the frequency of base i in site l for all 45 base sequences, p_i are probabilities of A, U, C and G bases in the RNA region with latent periodicity, and N is the number of periods.

Periods contain very diverged sites with almost uniform occurrence of all bases and of very conservative sites. For example, $\chi^2(13) = 11.3$ and 17 periods of this site contain 9 A and 8 G; $\chi^2(2) = 8.1$ and 17 periods contain 2 A, 12 U and 3 C.

A similar situation is observed for latently periodical RNA regions that are shown in Figure 1 and Table II. Table II shows the $2I'(n,\text{RNA})$ of the found regions with the latent periodicity for the transcriptional factor SP1 (HSTFSP1 clone) (Kadonaga *et al.*, 1987), human homeobox genes C1 (HSHOXB clone) (Simeone *et al.*, 1987) and C8 (HSHOX2I clone) (Acampora *et al.*, 1989), the ABL oncogene (HSABL clone) (Fainstein *et al.*, 1989), the FMS oncogene (HSCFMS clone) (Coussens *et al.*, 1986) and the epidermal growth factor receptor gene (HSEGFRS clone) (Merlino *et al.*, 1985). Latent periods of those genes have lengths of 87 up to 111 bases.

The latent periodicity is revealed for DNA-binding proteins. The transcription factor SP1 from HSTFSP1 clone, and the human C1 and C8 homeobox genes (HSHOXB and HSHSHOX2I clones), belong to this protein class (Gehring *et al.*, 1994). The GLI oncogene having DNA-binding activity (Kinzler *et al.*, 1988) also belongs to this protein type.

Some of the found regions with latent periodicity have fractional periodicity. It is observed for HSARAF1R and HSGLI clones. The period equal to 84 bases in the HSARAF1R clone (Figure 1B) has the included period equal to 21 bases. The region with latent periodicity in the HSGLI clone has a period equal to 117 bases and the included period equal to 39 bases. However, the HSRBS clone contains one sharp period equal to 45 bases. $X(n)$ spectra for regions embedded in Table I show that the periods found do not contain the included fractional periods.

The integrated periods for all found regions with latent periodicity are shown in Table III. The first base of the integrated period is the first codon base. It is extracted by enclosing the period in a ring and by a cyclic transposition of the bases. Information divergence $R_i(l)$ for each base i and for each site l of the period is the base for the integrated period determination. The composition information divergence is calculated as:

$$R_i(l) = f_i(l)\ln(f_i(l)/p_i) \qquad (8)$$

Here, $f_i(l) = n_i(l)/N$, and p_i and N are found using equation (7). The base which has $R_i(l) \geq 0.2$ is chosen for each site l. The choice of threshold significance R_i equal to 0.2 is based on the value $N \times R_i$ being less than 5.0. The sums of all $N \times R_i$ are distributed as χ^2 with three degrees of freedom (Kullback, 1959). If for two or more bases $R_i(l) \geq 0.2$, then these bases are shown one under the other in the integrated sequence.

Each site of the integrated sequence shows the bases creating latent periodicity. For example, the first base of the integrated sequence of the HSRBS clone is C. It means that 1, 46, 91, 136, ... bases of the region with latent periodicity of the HSRBS clone are enriched by C. This enrichment exceeds

Table II. The modified mutual information 2*I'* between DNA regions with latent periodicity and different artificial sequences

Period	HSTFSP1	HSHOXB	HSHOX2I	HSABL	HSCFMS	
	HSEGFRS					
3	5.6	26.9	15.9	93.6	117.5	26.9
6	11.5	30.0	18.5	92.8	112.8	19.3
9	10.9	20.7	16.1	95.4	140.8	33.8
12	7.2	31.6	26.0	99.2	107.5	15.6
15	28.9	27.7	0.4	83.0	120.2	22.0
18	7.9	14.5	12.3	99.5	138.1	36.8
21	−3.7	31.2	17.0	76.3	111.3	7.4
24	13.0	26.3	33.3	104.1	111.6	43.6
27	0.7	42.9	9.8	84.9	133.1	30.8
30	30.9	18.9	−9.5	79.6	114.1	−9.5
33	9.4	−3.2	5.1	88.7	124.9	22.9
36	1.0	14.5	14.5	100.2	140.5	32.1
39	23.2	23.2	−15.1	97.5	118.6	47.4
42	5.8	70.1	36.4	84.1	109.3	9.1
45	35.8	16.2	11.1	80.5	135.2	23.2
48	29.4	1.2	66.3	134.3	141.3	52.3
51	21.0	43.9	13.7	94.0	123.0	30.3
54	27.3	47.1	19.7	105.6	131.5	49.2
57	46.3	82.9	30.1	117.7	112.9	46.3
60	24.8	30.8	49.6	73.6	135.2	−13.4
63	23.3	46.4	29.5	77.6	177.5	27.4
66	21.8	28.1	34.5	125.6	138.4	21.8
69	41.8	59.9	44.1	107.9	73.9	39.6
72	49.6	70.6	58.8	117.6	165.6	90.0
75	43.6	94.2	16.7	125.1	146.5	52.9
78	51.5	37.4	6.0	149.1	122.0	56.3
81	15.1	92.5	28.8	77.2	191.8	62.2
84	51.0	102.0	78.5	126.2	107.3	38.8
87	39.5	82.4	46.9	237.9	117.2	79.8
90	65.5	4.1	45.2	116.3	141.6	25.5
93	45.9	79.7	66.5	104.1	164.3	40.8
96	81.0	41.5	94.6	160.6	169.6	72.9
99	52.6	57.9	47.3	101.6	264.7	71.3
102	88.9	86.1	58.7	137.9	165.1	75.0
105	170.4	73.3	35.4	119.0	221.3	65.0
108	85.6	108.8	52.1	169.4	157.0	100.0
111	83.6	190.7	177.9	146.5	184.3	181.1
114	108.6	186.3	76.3	135.9	126.7	90.8
117	77.2	106.9	45.8	137.6	195.3	89.0
120	66.5	78.2	105.1	114.3	187.6	96.1
123	67.3	109.5	97.2	176.4	193.0	70.2
126	74.0	142.4	92.2	142.4	225.9	129.6
129	51.0	118.3	68.8	193.8	207.5	68.8
132	78.8	81.9	103.7	202.7	172.1	103.7
135	70.4	104.8	76.6	163.7	201.3	89.0
138	96.3	99.5	86.8	158.6	112.4	86.8
141	97.3	103.8	150.1	201.8	234.0	133.3
144	91.8	131.3	114.7	196.7	196.7	141.4
147	51.2	109.2	115.8	132.6	177.3	109.2
150	103.6	116.9	83.8	179.0	186.1	70.9
153	–	124.6	114.6	155.9	205.9	142.0
156	–	115.7	56.0	218.7	150.3	105.5
159	–	79.6	96.3	224.4	231.9	113.3
162	–	124.8	83.6	153.0	207.6	131.8
165	–	115.4	77.6	140.0	198.2	140.0
168	–	166.5	116.4	199.8	166.5	126.9
171	–	113.8	135.2	193.9	247.5	160.6
174	–	–	114.8	281.1	143.6	150.9
177	–	–	101.4	212.4	208.5	90.8
180	–	–	142.2	183.4	241.4	95.1

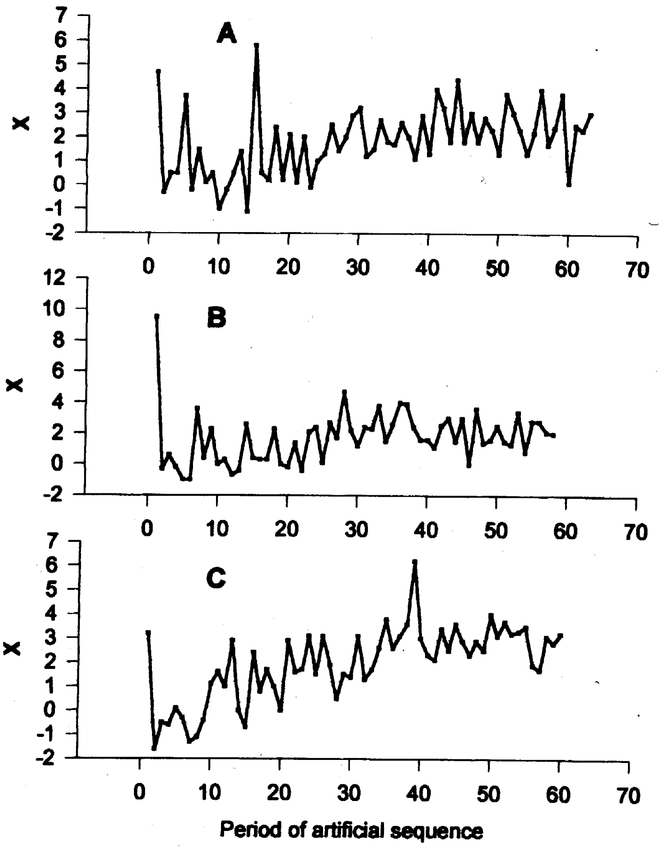


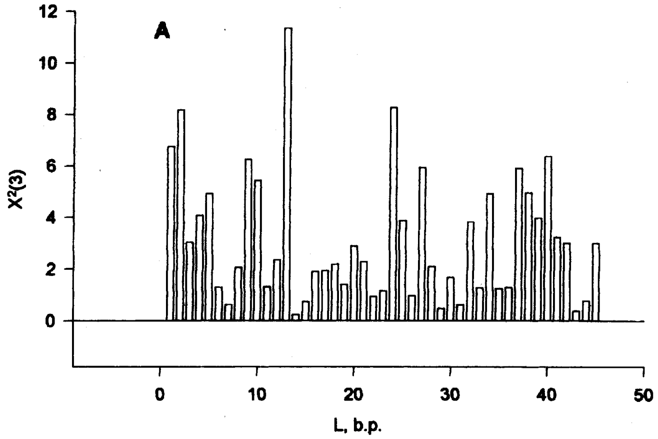
Fig. 2. The *X*(*n*) spectra for regions with latent periodicity from Figure 1. The length of period is shown as divisible by three. (A) mRNA region from HSRBS clone. (B) mRNA region from HSARAF1R clone. (C) mRNA region from HSGLI clone.

the average frequency C for the region with latent periodicity.

The method of searching for a periodicity of a symbolic sequence has been developed earlier (McLachlan, 1977b). The method is based on writing out the sequence in rows of *N* columns to detect a period of *N*. Then the symbols of the sequence are combined in two groups and a two-letter alphabet is introduced (A and B letters). The amalgamation of symbols in the two groups is performed on the bases of some common characteristics of the symbols. For amino acid sequence, those qualities may be hydrophobicity of amino acids, charges of amino acids and some other qualities (McLachlan, 1977b). However, the introduction of two letter alphabet limits the types of periodic sequences that may be found. For example, if we have the period {(a/t)(t/g/c)(c/a)}, then a new alphabet A = {a,t}, B = {c,g} permits a period to be revealed in the first column, but periodicity is not revealed in the second and third columns. If we introduce a new alphabet A = {t,g,c}, B = {a}, then we may find a periodicity of the second column only. An analysis of periodicity by the introduction of a new two-letter alphabet is effective if there is a similar distribution of the base types in the columns.

Table III. Integrated periods of the regions with latent periodicity, n is an arbitrary base

```
HSRBS
10         20        30       40
CTTAAAnCTGnAGnnnAnnnTnnTGnCnnAnTnAnAGAACATnnn
            A                          A
HSRAF1R
         10        20        30        40        50        60
GTACATGGnCnACAnTGACnTCCAGATGTTCAACCCCATCGnnAACAGGGCCCAGnTCAT
   G    C                        GA    A         T       T
         70        80
CGTCATTGCnnAGCGGTTTGAGCT
  A  A              T
HSGLI
         10        20        30        40       50         60
TCTCnTTCCCAnAACAGTnCCAAAnnnnACnTnGGTGCTCATCTTGnGGnGnnCCATGGG
   G          G     G  G                G
         70        80        90        100        110
GAAnCTCAACnGGCCnCATATnGTCCTCGnGGnGGACCGGAATTTACCCCTCnCnAn
        T
HSTFSP1
         10        20        30        40        50        60
GCCAATTnCCAAGAGACTCTTAATnAGCCCAACATCTTGGnAATTnTnAACTCTGTTnCC
              A             TT  T      C
         70        80        90        100
ACCTTnnTATCAGCACCAAnCTCTGAGGGnGnCTTTATCAnGAAG
                                G          G
HSHOXB
         10        20        30        40        50        60
nAGAAAnACTGAAATCACCTCTTATCnCCTCACnCGnAGGTCGnCTTCGGTTCCACCCTT
     C      CC                 T     G                G
        70        80        90         100        110
TGTCnTnnnGAGnGnAAnTTCTAGACCTAGGACCAGAAnGAAGnGATGAAG
          C        T C      TT          C     A
HSHOX2I
         10        20        30        40        50        60
nTGAAnTTCAACAGCTACGnGACCnAGGnnnGnnnCGTnTCnGnnGTCTACnCAACnGAG
 A               C   A                              CC
         70        80        90      100         110
GGCCCTnGTAAAnACTTCTATCCnnCGCAnCAGACCnnAnAnGnAAAGnAG
     A   C   C   AACC
HSABL
         10        20        30        40        50        60
GnnnnCATGAAGnTGTnCATnnAGnAnAAnnTCnnnGAGCnnGAnTnCTTCnTTGTGGnC
                          G                  G
         70        80
nnTGAnATCTTTnAAGnGAACnnCCAG
       G   G        G   TT
HSCFMS
         10        20        30        40        50        60
nGCnAGCnTnTGnACnnTGnTCACGTnnnGAAGnTnGTnGACTnCnnnTTCAATATCCTC
       C                    C   C                        A
         70       80         90
nTCnACAAGnAnTACCTCnACAAGGTCGTTGGCATGCTC
               A         A  A     A
HSEGFRS
         10        20       30        40        50       60
ATGTTGCATGnTnnnTATTCnATTGAnGGCTnCTnnnTGTTTGnGACnTGCCTTGnAGnn
  T                A        G
        70        80        90        100        110
TGCTTnnnnAATGnCATCAGCGTnGnCTAnAACTTCAnAGGCGCAGGCCAC
  T                     A                 GCAA  A
```



B

ACUUCCAGAGGUUGAAAAUCUUUCUAAACGAUACGAAGAAAUUUA
UUUUAUACCUUUUAUGAAUUCUCUUGGACUUGUAACAUCUAAUGG
UCUUAAAAAUAAAGAUCUAGAUGCAAGAUUAUUUUUGGAUCAUGA
UAAAACUCUUCAGACUGAUUCUAUAGACAGUUUUGAAACACAGAG
AACACCACGAAAAAGUAACCUUGAUGAAGAGGUGAAUGUAAUUCC
UCCACACACUCCAGUUAGGACUGUUAUGAACACUAUCCAACAAUU
AAUGAUGAUUUUAAAUUCAGCAAGUGAUCAACCUUCAGAAAAUCU
GAUUUCCUAUUUUAACAACUGCACAGUGAAUCCAAAAGAAAGUAU
ACUGAAAAGAGUGAAGGAUAUAGGAUACAUCUUUAAAGAGAAAUU
UGCUAAAGCUGUGGGACAGGGUUGUGUCGAAAUUGGAUCACAGCG
AUACAAACUUGGAGUUCGCUUGUAUUACCGAGUAAUGGAAUCCAU
GCUUAAAUCAGAAGAAGAACGAUUAUCCAUUCAAAAUUUUAGCAA
ACUUCUGAAUGACAACAUUUUUCAUAUGUCUUUAUUGGCGUGCGC
UCUUGAGGUUGUAAUGGCCACAUAUAGCAGAAGUACAUCUCAGAA
UCUUGAUUCUGGAACAGAUUUGUCUUUCCCAUGGAUUCUGAAUGU
GCUUAAUUUAAAAGCCUUUGAUUUUUACAAAGUGAUCGAAAGUUU
UAUCAAAGCAGAAGGCAACUUGACAAGAGAAAUGAUAAAACAU

Fig. 3. The comparison of 45 base periodical sequences from mRNA region with latent periodicity of HSRBS clone (Lee *et al.*, 1987). (**A**) The $\chi^2(3)$ for all 45 positions is shown. The probability $1 - (1 - \alpha)^{45}$ is <0.05 if χ^2 is >20. The probability α is equal to $P(\chi^2 \geq 20)$. (**B**) All the 17 periods from regions with latent periodicity of HSRBS clone. The 45 base long sequences are shown as each under the other.

Moreover, if we consider triplet periodicity, there are many types of junctions of triplets in two letters and it is very hard to test all types of junctions.

The mathematical method developed is easily applicable to DNA sequence analysis. It is sufficient to change U into T only. The regions of DNA sequences of some genes with latent periodicity were found earlier (Korotkov and Korotkova, 1995). Results of this work and earlier work show that no less than 30% of human genes from the EMBL data bank have regions with latent periodicity. Latent periodicity is not similar to a homological alternation of DNA bases. It is possible to assume that latent periodicity is a typical sign of some genes and can be related to the evolutionary origin of genes by a process of multiple duplications. Then latent periodicity is a reflection of ancient evolutionary events in gene sequences.

It is also possible to assume that the latent periodicity can also have a certain function in a cell. The latent periodicity can provide certain bends in DNA coding regions (McNamara *et al.*, 1990; Carrera and Azorin, 1994; Goodsell and Dickerson, 1994).

Acknowledgement

This work was carried out with the support of the Russian Fund of Fundamental Investigation.

References

Acampora,D., D'Esposito,M., Faiella,A., Pannese,M., Migliaccio,E. Morelli,F., Stornaiuolo,A., Nigro,V., Simeone,A. and Boncinelli,E. (1989) The human hox gene family. *Nucleic Acids Res.*, **17**, 10385–10409.

Beck,T.W., Huleihel,M., Bonner,T.I. and Rapp,U.R. (1987) The complete coding sequence of the human A-RAF-1 oncogene and transforming activity of a human A-RAF carrying retrovirus. *Nucleic Acids Res.*, **15**, 595–609.

Bina,M. (1994) Periodicity of dinucleotides in nucleosomes derived from Simian virus 40 chromatin. *J. Mol. Biol.*, **235**, 198–208.

Bliskovsky,V.V. (1991) Tandem DNA repeats in vertebrate genomes: structure, possible mechanisms of creation and evolution. *Mol. Biol. (Russian)*, **25**, 965–982.

Borstnik,B., Pampernik,D., Lukman,D., Ugarkovic,D. and Plohl,M. (1994) Tandemly repeated pentanucleotides in DNA sequences of eukaryotes. *Nucleic Acids Res.*, **22**, 3412–3417.

Carrera,P. and Azorin,F. (1994) Structural characterization of intrinsically curved AT-rich DNA sequences. *Nucleic Acids Res.*, **22**, 3671–3680.

Chechetkin,V.R., Knizhnikova,L.A. and Turygin,A.Yu. (1994) Three-quasiperiodicity, mutual correlations, ordering and long modulations in genomic nucleotide sequences of viruses. *J. Biomol. Struct. Dynam.*, **12**, 271–299.

Cheever,E.A., Overton,G.C. and Searls,B.B. (1991) Fast fourier transform-based correlations of DNA sequences using complex plane encoding. *Comput. Applic. Biosci.*, **7**, 143–154.

Coussens,L., Van Beveren,C., Smith,D., Chen,E., Mitchell,R.L., Isacke,C.M.,Verma,I.M,. and Ullrich,A. (1986) Structural alternation of viral homologue of receptor proto-oncogene FMS at carboxyl terminus. *Nature*, **320**, 277–280.

Fainstein,E., Einat,M., Gokkel,E., Marcelle,C., Croce,C.M., Gale,R.P and Canaani,E. (1989) Nucleotide sequence analysis of human ABL and BCR-ABL cDNA's. *Oncogene*, **4**, 1477–1481.

Gehring,W.J., Qian,Y.Q., Billeter,M., Furucubo-Tokunaga,K., Schier,A.F., Resendez-Perez,D., Affolter,M., Otting,G. and Wuthrich,K. (1994) Homeodomain-DNA recognition. *Cell*, **78**, 211–223.

Goodsell,D.S. and Dickerson,R.E. (1994) Bending and curvature calculations in B+DNA. *Nucleic Acids Res.*, **22**, 5497–5503.

Handbook of Applicable Mathematics (1984) Wiley-Interscience, John Wiley & Sons, New York, Vol. IV.

Hearne,C.M., Ghosh,S. and Todd,J.A. (1992) Microsatellites for linkage analysis of genetic traits. *Trends Genet.*, **8**, 288–294.

Kadonaga,J.T., Carner,K.R., Masiarz,F.R. and Tjian,R. (1987) Isolation of cDNA encoding transcription factor SP1 and functional analysis of the DNA binding domain. *Cell*, **51**, 1079–1090.

Kinzler,K.W., Ruppert,J.M., Bigner,S.H. and Vogelstein,B. (1988) The GLI gene is a member of the kruppel family of zinc finger proteins. *Nature*, **332**, 371–374.

Korotkov,E.V. (1994) Fast method of homology and purine-pyrimidine mutual relations between DNA sequences search. *DNA Sequence*, **4**, 413–415.

Korotkov,E.V. and Korotkova,M.A. (1993) Enlarged similarity of nucleic acids sequences. Preprint of Moscow Physical Engineering Institute, N.012-93, Moscow.

Korotkov,E.V. and Korotkova,M.A. (1995) Latent periodicity of DNA sequences of some human genes. *DNA Sequence*, **5**, 353–358.

Kullback,S. (1959) *Information Theory and Statistics*. John Wiley & Sons, New York.

Lee,W.H., Shew,J.Y., Hong,F.D., Sery,T.W., Donoso,L.A., Young,L.J., Bookstein,R. and Lee,E.Y.H.P. (1987) The retinoblastoma susceptibility gene encodes a nuclear phosphoprotein associated with DNA binding activity. *Nature*, **329**, 642–645.

Makeev,V.Ju. and Tumanyan,V.G. (1994) On a link between the distance and correlation analysis of various types and Fourier transformation used for the search of the periodical patterns in the primary structures of the biopolymers. *Biophysics (Russian)*, **39**, 294–297.

McLachlan,A.D. (1977a) Repeated helical pattern in apolipoprotein A-I. *Nature*, **267**, 465–466.

McLachlan,A.D. (1977b) Analysis of periodic patterns in amino acid sequences: collagen. *Biopolymers*, **16**, 1271–1297.

McLachlan,A.D. (1993) Multichannel Fourier analysis of patterns in protein sequences. *J. Phys. Chem.*, **97**, 3000–3006.

McLachlan,A.D and Stewart,M. (1994) Analysis of repeated motif in the talin rod. *J. Mol. Biol.*, **235**, 1278–1290.

McNamara,P.T., Bolshoy,A., Trifonov,E.N. and Harrington,R.E. (1990) Sequence-dependent kinks in curved DNA. *J. Biomol. Struct. Dynam.*, **3**, 529–538.

Merlino,G.T., Ishii,S., Whang-Peng,J., Knutsen,T., Xu,Y.H., Clark,A.J.L., Stratton,R.H., Wilson,R.K., Ma,D.P., Roe,B.A., Hunts,J.H., Shimizu,N. and Pastan,I. (1985) Structure and localization of genes encoding aberrant and normal epidermal growth factor receptor RNAs from A431 human carcinoma cells. *Mol. Cell. Biol.*, **5**, 1722–1734.

Pizzi,E., Linni,S. and Frontali,C. (1990) Detection of latent sequence periodicities. *Nucleic Acids Res.*, **18**, 3745–3752.

Silverman,B.D. and Linsker,R. (1986) A measure of DNA periodicity. *J. Theor. Biol.*, **118**, 295–300.

Simeone,A., Mavilio,F., Acampora,D., Giampaolo,A., Faiella,A., Zappavigna,V., D'Esposito,M., Pannese,M., Russo,G., Boncinelli,E. and Peschle,C. (1987) Two human homeobox genes, C1 and C8: structure analysis and expression in embryonic development. *Proc. Natl Acad. Sci. USA*, **84**, 4914–4918.

Stewart,M. and McLachlan,A.D.(1975) Fourteen actin-binding sites on tropomyosin? *Nature*, **257**, 331–333.

Todd,J.A. (1992) La carte des microsatellites est arrivee. *Hum. Mol. Genet.*, **1**, 663–666.

von Heijne,G. (1987) *Sequence Analysis in Molecular Biology*. Academic Press, San Diego, CA.

Voss,R.F. (1992) Evolution of long-range fractal correlations and **1/f** noise in DNA base sequences. *Phys. Rev. Lett.*, **68**, 3805–3808.

Yaglom,A.M. and Yaglom,I.M. (1960) *Probability and Information*. Nauka Press, Moscow.

Received on June 5, 1996; accepted on July 25, 1996