DNA Sequence—The Journal of Sequencing and Mapping, Vol. 5, pp. 353–358
Reprints available directly from the publisher
Photocopying permitted by license only

Latent periodicity of DNA sequences from some human gene regions

E. V. KOROTKOV and M. A. KOROTKOVA

N. N. Semenov Institute of Chemical Physics, 117977, Moscow, Russia and Department of Cybernetics, Moscow Physical Engineering Institute, 115409, Kashirskoe chosse, 31, Moscow, Russia

The mutual information is used to reveal of DNA sequences latent periodicity. Latent periodicity of DNA sequence is periodicity with low level of homology between any two periods inside DNA sequence. The mutual information between artificial numerical sequence and DNA sequence is calculated. The length of artificial sequence period is changed from 2 to 250. High level of mutual information between artificial and DNA sequences allows to find any type of latent periodicity of DNA sequence. The latent periodicity of some DNA coding regions is considered. For example, 24 exon of Apo B-100 gene from HSAPB21 clone contains latent period 84 bases long. The IGF-I receptor gene from HSIGFIRR clone contains the region with latent period 57 bases long. Possible significance of latent periodicity is discussed.

KEY WORDS: Sequence, computer analysis, periodicity, mutual information

The investigation of DNA sequence periodicity is the way of understanding of the structure of human and other genomes. The periodical sequences have been found in human and other genomes as satellite and minisatellite sequences (Bliskovsky, 1991). Those sequences contain short DNA fragments that are repeated by often. The amino acid sequences of some proteins may be considered as periodical sequences also (McLachlan, 1993). Mathematical methods of a symbolic sequence analysis are developing in present time (Voss, 1992; McLachlan, 1993; Mclachlan and Stewart, 1994). The powerful mathematical methods of periodicity analysis of numerical sequences are applied to investigation of symbolic sequence periodicity. The transformation of a symbolic sequence to a numerical sequence is used often for such analysis. Also, the weights of letter pairs of a symbolic sequence are introduced by different ways (Silverman and Linsker, 1986; Mclachlan and Stewart, 1994). One way of such weights introducing is the next: $x_i x_j$ is equal to 1 if $x_i = x_j$ and $x_i x_j$ is equal to 0 if $x_i \neq x_j$ (Voss, 1992). Those investigations have shown a periodicity of some amino acid sequences. However, the investigation of DNA sequence periodicity is being focused on the search of homological periodicity (including incomplete homological periodicity) (Makeev and Tumanyan, 1994). But there is some DNA sequence periodicity that is different from homological periodicity. For example, the sequence may have the period (A/G) (G/C/T)N(C/G) (A/T) (G/T) (G/C/A). Here N is the arbitrary base. The first base of this period may be A or G, the second base may be G, C or T etc. Such DNA base periodicity of sequence may be statistically important if the sequence has enough length. Search of such latent periodicity is difficult if known mathematical methods are applied. The mathematical method for revealing of latent DNA sequence periodicity is developing in the present report. This method uses the principle of the enlarged similarity of symbolic sequences (Korotkov and Korotkova, 1993). Latent periodicity of regions of apolipoprotein B-100 gene, apolipoprotein E gene, tropomyosin gene, IGF-I receptor gene, 5-lipoxygenase gene, methylmalonyl CoA mutase gene, lamin B2 gene and hematopoietic lineage cell protein are shown.

To reveal the latent periodicity of DNA sequence is used the comparison of artificial periodical sequence with a DNA sequence. The alphabet of the artificial sequence contains the S_i letters. Artificial sequence $S_1S_2S_1S_2S_1S_2 \ldots$ is created if the period equal to the two bases is analyzed in a DNA sequence. The sequence $S_1S_2 \ldots S_nS_1S_2 \ldots S_nS_1S_2 \ldots S_n \ldots$ is created if period equal to the n bases in DNA sequence is analyzed. Length of artificial

Address for correspondence: Dr. Eugene Korotkov, N. N. Semenov Institute of Chemical Physics, 117977 Moscow, Russia. Fax: 7-095-938-2156; E.mail: chembio@glas.apc.org.

sequence is equal to the length of analyzed DNA sequence.

Artificial sequences with periods from 2 to 250 are compared with analyzed DNA sequence by one to another. Measure of sequences similarity is mutual information. The M matrix is filled for mutual information calculation. Matrix M elements are the numbers that show numbers of each type coincidence between artificial sequence and DNA sequence. The matrix M dimension is n × 4. Signs of M matrix lines are bases A, T, C and G of the DNA sequence. Signs of M matrix columns are S_i letters of the artificial sequence. The mutual information is calculated using the formula (Kullback, 1959):

$$I = \sum_{1}^{4}\sum_{1}^{n} m_{ij} \ln m_{ij} - \sum_{1}^{4} x_i \ln x_i - \sum_{1}^{n} y_j \ln y_j + L \ln L \quad (1)$$

Here m_{ij} are a matrix M elements; x_i are numbers of A, T, C and G bases in DNA sequence; y_j are numbers of S_j symbols in artificial sequence; L is a length of sequences. The 2I is distributed as X^2 with 3 (n-1) degree of freedom. It allows to estimate the probability of accidental origin of any period in a DNA sequence.

All periods may be classified on two classes. The first class includes the periods that lengths are the simple numbers. The second class includes periods that lengths are equal to the composite numbers. Let we have the composite δ period that equals to βγ. Let the β period have lowest probability $\alpha = P(X^2 \geq 2I)$ for all being analyzed periods of the DNA sequence. It was shown that (Yaglom and Yaglom, 1960):

$$I(\gamma, DNA) + I(\beta, DNA) \leq I(\beta\gamma, DNA) \quad (2)$$

Here and below I(x, DNA) means a mutual information between artificial sequence with period length equal to x and DNA sequence. The formula (2) also means that the mutual information I(kβ, DNA) may be equal or may be more than I(β, DNA) if k is increased. For example, if the mutual information is calculated between artificial sequences with lengths of periods 3,6,9,12, . . . and DNA than the mutual information for 6,9,12, . . . periods are equal or more I(3, DNA). The calculation of probability $F(X^2 \geq 2I(\delta, DNA))$ at condition $2I(\delta, DNA) \geq 2I(\beta, DNA) + 2I(\gamma, DNA)$ for the appraisal of accidental formation probability of the δ period is required. The probability F is possible to consider as probability of accidental formation of period equal to δ in DNA sequence.

The DNA sequence regions are considered in this report that have the main period length is equal to 3 bases and α probability is less than 10^{-5}. Moreover, those sequences have the minor period (multiple to 3 bases) with probability F less than 10^{-4}.

The regions with latent periodicity are found in the human clones from EMBL data bank. The artificial sequence containing 1000 letters is compared with first 1000 bases of analyzed clone. Clone isn't analyzed if long of clone is less than 1000 bases. The region of DNA that has $\alpha = P(X^2 \geq 2I(n, DNA))$ minimum inside 1000 bases is found by way of independent variation of a left and right boundaries. Those variations of boundaries are done for all artificial sequences with n from 2 to 250. If a minimum α is less or equal to 10^{-5} than the F probability of the minor periods is calculated also. The minor period with minimum F is chosen as latent period. The 500 bases shift of the artificial sequence is executed after such analysis and artificial sequence is compared with the region of DNA sequence from the base number 501 to the base number 1500. Then this analysis is repeated. If the probability α of main period is more than 10^{-5} than a value of the shift is equal to 100 bases and the analysis is repeated. Full length of human DNA clone is analyzed. The all human clones from EMBL data bank have been analyzed by such way.

The analysis has found various DNA sequences with minor or latent periodicity. The 9 such regions with lowest F values from different human genes are shown in Figure 1 and Table 1. Values 2I' (n, DNA) = 2I (n, DNA) −3 (n-1) are shown in Figure 1 and Table 2. The 3(n-1) is a middle mutual information between accidental artificial sequence that the alphabet contains n letters and accidental sequence that the alphabet contains 4 letters (Kullback, 1959).

The main period is equal to 3 bases and the values of 2I(3, DNA) are from 38 to 150 for all 9 regions. It corresponds to probability α from 10^{-6} to less than 10^{-10}. Period equal to 3 bases has been found in coding regions. This period is the trait of coding regions. The minor periods multiple to 3 bases are found in those sequences. The 2I' values for different minor periods are shown in Figure 1 for 3 DNA regions from HSAPB21 (Ludwig *et al.*, 1987), HSTRO (Lin and Leavitt, 1988) and HSAPOE4 (Das *et al.*, 1985) clones. Those clones contain apolipoprotein B-100, tropomyosin and apolipoprotein E genes. Period equal to 84 bases are found in 24 exon of Apo B-100 gene. The same length of period has been found in DNA sequences of a zinc finger genes also. Period

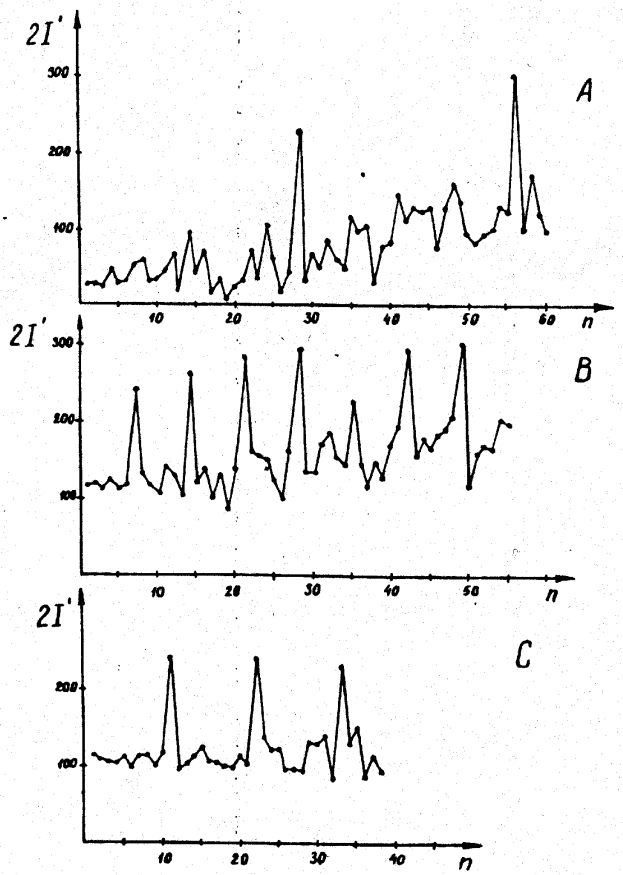


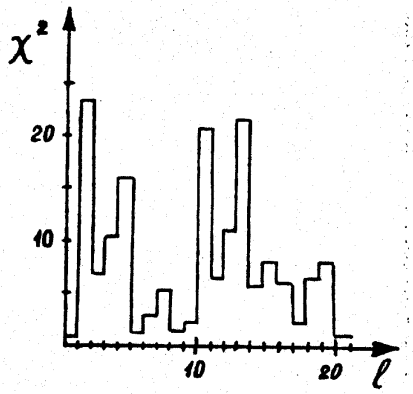
Figure 1 The dependence 2I' upon period of artificial sequence. A - DNA region (3586–4475 bases) from HSAPB21 clone; B - DNA region (3784–4307 bases) from HSTRO clone; C - DNA region (244–994 bases) from HSAPOE4 clone.

Table 1 The coordinates of regions with latent periodicity in EMBL clones.

Clone From EMBL Data Bank	*Genes Where the Latent Periodical Sequences Are Found*	*Coordinates of Found Sequences in EMBL Clone*	*Period Length (Bases)*
1. HSAPB21	APOLIPOPROTEIN B-100, EXON 29	13636–14224	33
2. HSHEAM	HEAMATOPOIETIC LINEAGE CELL PROTEIN	30–993	111
3. HSIGFIRR	INSULIN-LIKE GROWTH FACTOR I RECEPTOR	3365–4093	57
4. HSLOX5A	5-LIPOXYGENASE	34–1000	15
5. HSMCM	METHYLMALONYL COA MUTASE	1146–2086	115
6. HSLAMB2	LAMININ B2	4107–4906	21

equal to 21 bases is revealed in the fragment of tropomyosin gene and period equal to 33 bases is revealed in the fragment Apo E gene.

It is important to note that those minor periodicity of the 9 found regions is difficult to find using the homology between the sequences and those periods may be considered as latent. Level of the homology between any periods of the Apo B-100 24 exon and such homology level between any periods of the tropomiosin gene and between any periods of the Apo E4 gene are not statistically important. The periods from HSTRO clone are shown in Figure 2. The



A

B

GCCGCTCCTGCTGCAGCCCCA
GGGCCCCTCGCCGCCGCCACC
ATGGACGCCATCAAGAAGAAG
ATGCAGATGCTGAAGCTCGAC
AAGGAGAACGCCTTGGATCGA
GCTGAGCAGGCGGAGGCCGAC
AAGAAGGCGGCGGAAGACAGG
AGCAAGCAGCTGGAAGATGAG
CTGGTGTCACTGCAAAAGAAA
CTCAAGGGCACCGAAGATGAA
CTGGACAAATACTCTGAGGCT
CTCAAAGATGCCCAGGAGAAG
CTGGAGCTGGCAGAGAAAAAG
GCCACCGATGCTGAAGCCGAC
GTAGCTTCTCTGAACAGACGC
ATCCAGCTGGTTGAGGAAGAG
TTGGATCGTGCCCAGGAGCGT
CTGGCAACAGCTTTGCAGAAG
CTGGAGGAAGCTGAGAAGGCA
GCAGATGAGAGTGAGAGAGGC
ATGAAAGTCATTGAGAGTCGA
GCCCAAAAAGATGAAGAAAAA
ATGGAAATTCAGGAGATCCAA
CTGAAAGAGGCAAAGCACATT
GCTGAAGATGCCGACCGCAAA
TATGAAGAGGTGGCCCGTAAG
CTGGTCATCATTGAGAGCGAC
CTGGAACGTGCAGAGGAGCGG
GCTGAGCTCTCAGAAGGCCAA
GTCCGACAGCTGGAAGAACAA
TTAAGAATAATGGATCAGACC
TTGAAAGCATTAATGGCTGCA
GAGGATAAGTACTCGCAGAAG
GAAGACAGATATGAGGAAGAG
ATCAAGGTCCTTTCCGACAAG
CTGAAGGAGGCTGAG

Figure 2 The comparison of 21 base periodical sequences from DNA region with latent periodicity of HSTRO clone. A. The X^2 for all 21 positions is shown. The probability $1-(1-a)^{21}$ is less than 0.05 if X^2 is more than 18. The probability a is equal $P(X^2 \geq 18)$). B. The all 36 periods from regions with latent periodicity of HSTRO clone. The 21 base long sequences are shown as each under other.

values of $X^2(l)$ are shown under the sequence. The $X^2(l)$ shows the difference between middle frequencies of A, T, C, and G bases for all periods and the frequencies of A, T, C and G bases for site l.

$$X^2(\ell) = \sum_{1}^{4} (n_i(\ell) - p_i N) / p_i N \qquad (3)$$

Here the $n_i(l)$ are the numbers of A, T, C and G bases in the l site for all 21 nucleotide sequences; p_i are the frequencies of A, T, C and G bases for the region with latent periodicity; N is the number of periods in the region with latent periodicity. The $X^2(l)$ is distributed with 3 degrees of freedom. Figure 2 shows that the periods have as very turbid sites as conservative sites. For example, $X^2(2)$ is equal to 23.3 and all 36 periods of this site contain 5A, 22T, 7C and 2G; $X^2(11)$ is equal to 18.5 and all 36 periods contain 5A, 13T, 17C and 1G. The similar situation is for all sequences with the latent periodicity that are shown in Figure 1 and Table 2. The 21′ (n, DNA) values for the regions of exon 29 of Apo B-100 gene (HSAPB21 clone), IGF-I receptor gene (HSIGFIRR clone) (Ullrich *et al.*, 1986), 5-lipoxygenase (HSLOX5A clone) (Matsumoto *et al.*, 1988), methylmalonyl CoA mutase gene (HSMCM clone) (Jansen *et al.*, 1989), lamin B2 gene (HSLAMB2 clone) (Pikkarainen *et al.*, 1988) and hematopoietic lineage cell protein (HSHEAM clone) (Kitamura *et al.*, 1989) are shown in Table 2.

It should be noted that those regions obtain the latent periodicity of different length (from 15 to 111 bases). Latent periodicity of the HSIGFIRR clone is especially important because this region contains the tyrosine kinase domain. The domain is very evolutionary conservative and is found for other receptors also (Ullrich *et al.*, 1986).

The present report shows that some DNA sequences of human genes contain the regions with latent periodicity. Such periodicity is different from homology similarity of DNA bases. It may be proposed that some found minor periods reflect the evolutionary origin of genes by process of the multitude DNA tandem duplications. Also, it may be supposed that some DNA regions with latent periods reflect the principles of spatial organization of being coded proteins.

The developing mathematical method allows to create the data bank of a regions with different type of latent periodicity.

(Received 10 September 1994)

Table 2 The modified mutual information 2I′ between DNA regions with latent periodicity and different length artificial sequences. Periods are shown as multiple to three bases.

Period	*HSAPB21*	*HSHEAM*	*HSIGFIRR*	*HSLOX5A*	*HSMCM*	*HSLAMB2*
1	34.6	65.0	82.7	143.9	103.8	125.5
2	40.0	62.9	77.3	139.6	101.8	129.2
3	30.7	57.4	81.9	146.3	100.1	133.6
4	30.4	57.6	73.0	144.9	102.5	133.8
5	27.7	61.8	81.4	162.2	116.6	123.8
6	42.2	53.4	74.6	138.1	96.1	150.0
7	45.3	65.0	88.1	136.2	109.4	262.8
8	40.6	59.1	63.9	139.2	96.7	137.1
9	36.8	47.6	69.1	143.5	86.7	119.0
10	43.5	70.6	77.8	185.7	116.1	118.1
11	142.1	43.8	100.4	148.7	106.4	144.2
12	32.1	58.0	70.9	138.3	110.4	158.5
13	23.2	54.7	71.6	122.9	105.8	133.9
14	68.1	86.2	84.1	135.8	120.2	267.7
15	23.2	72.4	99.5	173.5	101.7	156.4
16	38.8	50.4	74.6	155.8	111.3	127.3
17	70.3	70.3	98.5	135.4	135.4	121.2
18	94.2	80.9	76.5	153.6	76.5	122.0
19	48.4	80.7	210.2	157.2	124.9	144.6
20	37.0	60.4	82.7	214.7	110.7	135.2
21	55.2	59.6	98.7	156.2	148.4	271.9
22	186.5	17.7	115.6	151.5	154.1	154.1
23	46.3	90.6	100.4	146.4	138.5	141.2
24	54.2	51.9	75.4	149.2	117.6	171.1
25	101.8	84.2	89.2	188.3	154.7	163.0
26	49.2	75.8	95.9	146.3	154.6	124.7
27	64.7	79.7	72.2	148.8	110.8	102.9
28	63.3	76.0	112.6	134.5	140.0	285.4
29	64.4	98.3	95.6	165.2	117.2	165.2
30	97.1	108.0	138.7	219.3	116.3	176.6
31	33.3	77.1	132.2	182.2	143.7	85.1
32	57.0	91.9	122.7	187.9	154.7	145.8
33	184.3	60.5	156.8	178.1	147.9	165.9
34	108.8	103.0	114.5	186.7	180.5	212.1
35	98.7	107.3	151.9	176.6	257.9	292.1
36	108.8	153.9	111.7	182.0	129.5	185.2
37	116.1	331.4	177.9	190.7	162.1	181.1
38	102.6	99.6	206.1	180.0	157.7	145.2
39	83.1	113.0	109.9	218.6	175.7	178.9
40	111.2	161.5	84.1	252.1	177.7	158.3
41	58.5	153.6	186.3	173.1	186.3	112.5
42	92.2	136.0	132.8	201.9	195.1	275.4
43	115.1	93.2	140.7	217.9	163.6	137.5
44	227.1	94.3	189.0	192.4	209.7	172.1
45	79.7	130.6	163.7	240.2	191.0	157.0
46	125.4	141.9	138.6	172.2	179.1	155.3
47	103.8	123.4	143.3	198.3	187.8	110.3
48	148.1	75.8	124.6	225.3	189.6	236.2
49	129.2	109.2	173.8	202.1	220.1	329.6
50	—	151.0	154.5	252.5	189.6	193.2

REFERENCES

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

Das, H.K., Mcpherson, J., Bruns, G.A.P., Karathanasis, S.K., Breslow, J.L. (1985). Isolation, characterization, and mapping to chromosome 19 of the human apolipoprotein E gene. *J. Biol. Chem.* **260**, 6240–6247.

Jansen, R., Kalousek, F., Fenton, W.A., Rosenberg, L.E., Ledley, F.D. (1989). Cloning of full-length methylmalonyl-coa mutase from a cDNA library using the polymerase chain reaction. *Genomics* **4**, 198–205.

Kitamura, D., Kaneko, H., Miyagoe, Y., Ariyasu, T., Watanabe, T. (1989). Isolation and characterization of a novel human gene expressed specifically in the cells of hematopoietic lineage. *Nucl. Acids Res.* **17**, 9367–9379.

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

Kullback, S. (1959). Information theory and Statistics. (New-York, London: John Wiley & Sons, Inc. Press).

Lin, C.S., Leavitt, J. (1988). Cloning and characterization of a cDNA encoding transformation sensitive tropomyosin isoform 3 from tumorigenic human fibroblasts. *Mol. Cell. Biol.* **8**, 160–168.

Ludwig, E.H., Blackhart, B.D., Pierotti, V.R., Caiati, L., Fortier, C., Knott, T., Scott, J., Mahley, R.W., Levy-Wilson, B., Mccarthy, B.J. (1987). DNA sequence of the human apolipoprotein B gene. *DNA* **6**, 363–372.

Makeev, V. Ju., Tumanyan, V.G. (1994). On 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.

Matsumoto, T., Funk, C.D., Raadmark, O., Hoeoeg, J.O., Joernvall, H., Samuelsson, B. (1988). Molecular cloning and amino acid sequence of human 5-lipoxygenase. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 26–30.

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

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

Pikkarainen, T., Kallunki, T., Tryggvason, K. (1988). Human laminin b2 chain: comparison of the complete amino acid sequence with the b1 chain reveals variability in sequence homology between different structural domains. *J. Biol. Chem.* **263**, 6751–6758.

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

Ullrich, A., Gray, A., Tam, A.W., Yang-Feng, T., Tsubokawa, M., Collins, C., Henzel, W., Bon, T.L., Kathuria, S., Chen, E., Jakobs, S., Francke, U., Ramachandran, J., Fujita-Yamaguchi, Y. (1986). Insulin-like growth factor I receptor primary structure: comparison with insulin receptor suggests structural dereminants that define functional specificity. *EMBO J.* **5**, 2503–2512.

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

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