Gene 421 (2008) 52–60

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Classification analysis of triplet periodicity in protein-coding regions of genes

F.E. Frenkel *, E.V. Korotkov

Bioengineering Centre of RAS, Moscow, Russia

ARTICLE INFO

Article history:
Received 4 March 2008
Received in revised form 14 May 2008
Accepted 6 June 2008
Available online 11 June 2008

Received by M. Di Giulio

Keywords:
Triplet periodicity
Classification
Coding regions
Reading frame
Shift
Inversion

ABSTRACT

We introduce a new concept of triplet periodicity class (TPC) and a measure of similarity between such classes. We performed classification of 472288 triplet periodicity (TP) regions found in 578868 genes from 29th release of KEGG databank. Totally 2520 classes were obtained. They contain 94% of 472288 found cases of TP. For 92% of TP regions contained in classes the same linkage of TP to open reading frame (ORF) is observed. For 8% of TP cases we revealed a shift between ORF of a gene and ORF common for majority of genes contained in a TPC. For these 8% of periodic regions the hypothetical amino acid sequences corresponding to ORF built by TPC were made. BLAST program has shown that 2679 hypothetical amino acid sequences have statistically significant similarity with proteins from UniProt databank. We suppose that 8% of TP regions contained in classes possess a mutation originating from ORF shift. Obtained TPCs can be used for identification of genes' coding regions as well as for searching for mutations arisen arising from ORF shift.

1. Introduction

Triplet organization of protein-coding DNA sequences is a property of all currently known living systems (Fickett, 1998; Staden, 1994; Baxevanis, 2001; Gutiérrez et al., 1994; Gao et al., 2005; Yin and Yau, 2007; Eskesen et al., 2004; Bibb et al., 1984; Konopka, 1994). The reason for this lies not only in structure of genetic code (Trifonov, 1999; Eigen and Winkler-Oswatitsch, 1981), but also in usage of "favorite" triplets of DNA residues for coding certain amino acids (Zoltowski, 2007; Antezana and Kreitman, 1999) and in saturation of proteins by particular amino acids (Karlin and Bucher, 1992; Zhang, 2005). It is also suggested that TP appeared as a result of necessity of mutation control via ORF shift (Trifonov, 1987). Investigation of gene TP can be a subject of interest for development of more powerful algorithms for searching DNA coding regions (Fickett, 1996), and also for analysis of DNA coding regions' evolution (Trifonov, 1999; Eigen and Winkler-Oswatitsch, 1981). For revealing TP, there are currently developed methods that use regularity of preferences for symbols in different triplet positions in DNA sequences. As a mathematical tool, they use Fourier transform (Issac et al., 2002; Makeev and Tumanyan, 1996; Tiwari et al., 1997), hidden Markov chains (Azad and Borodovsky, 2004; Henderson et al., 1997), neural networks (Snyder and Stormo, 1993; Thomas and Skolnick, 1994) and some other statistical methods based on position-dependent preferences for nucleotides in coding sequences (Fickett, 1996).

Certain problems arise when the coding potentials based on current mathematical methods are used. First, Fourier transform-based methods do not allow revealing TP with symbol insertions and deletions, and also do not allow distinguishing TP found in one DNA sequence from TP found in another one. This can be done by introducing a term of "triplet periodicity class" (TPC). Classes combine genes having closely-related TP. We consider the class to be a property that allows distinguishing TP found in one DNA sequence from TP found in another DNA region in quantitative manner. Methods based on dynamic programming allow detecting periodicity with insertions and deletions, but do now allow finding latent periodicity in DNA coding regions through using weight matrices for nucleotide pairs (Korotkov et al., 2003a,b). When using neural networks, learning sample of coding sequences (Snyder and Stormo, 1993; Thomas and Skolnick, 1994) is created. But possible antagonistic triplet periodicities will decrease specificity of TP for revealing coding regions.

TP is important for development of mathematical methods for DNA coding regions' prediction (Fickett, 1996). The task is to develop such a method of revealing TP that will detect it at higher statistical significance level. This will allow increasing specificity of DNA coding regions' detection. However, there are several problems on this way. First, mutational alterations of DNA sequences include not only

Abbreviations: CCM, central class matrix; ORF, open reading frame; TP, triplet periodicity; TPC, triplet periodicity class; TPM, triplet periodicity matrixI; TPC–ORF, reading frame of triplet periodicity class.

* Corresponding author. 60-letiya Oktyabrya prosp., 7/1, Moscow 117312, Russia. Tel.: +7 499 135 2161; fax: +7 499 135 0571.
E-mail address: felix.frenkel@gmail.com (F.E. Frenkel).

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.06.012

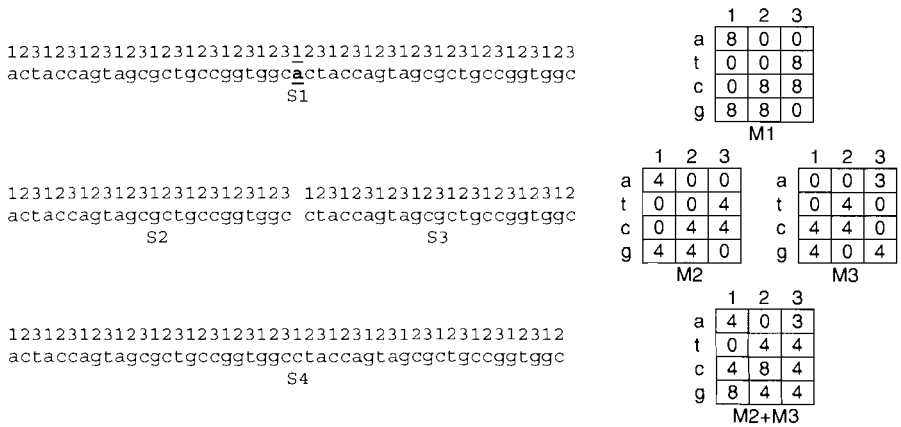

Fig. 1. Impact of one base deletion on TP of nucleotide sequence. Numerals over sequences S1, S2, S3 and S4 indicate nucleotide positions in ORF. The 25th base (underlined) was deleted from sequence S1. Consequently sequence S1 can be represented as two sequences S2 and S3. In these sequences there will be the same periodicity (matrices M2 and M3) but in sequence S3 it will be shifted by one base relative to new ORF appeared after deletion. This means that 1st column of matrix M2 corresponds to 3rd column of matrix M3, 2nd column of matrix M2 corresponds to 1st column of matrix M3, 3rd column of matrix M2 corresponds to 2nd column of matrix M3. Consequently we obtain sequence S4 that has TP matrix M4=M1+M2. After summation 1st column of matrix M2 is joined with 1st column of matrix M3 and so on. This results in joining of nonidentical columns and considerably decreases statistical significance of TP in sequence S4.

substitutions, but also symbols' insertions and deletions. They decrease the possibility of detecting TP in coding sequences of DNA bases (Fig. 1).

Second, different DNA sequences may contain antagonistic types of TP (Fig. 2). Such types, upon their concatenation into one learning sample (e.g., for neural network tuning), have lower specificity for coding region detection. We can avoid these problems and increase power of TP as coding potential if we single out classes of TP which combine genes possessing similar type of periodicity. Search for DNA coding regions can be carried out by methods of profile analysis (Gribskov and Veretnik, 1996) using obtained classes of TP as profiles. It is possible to search for TP with insertions and deletions by method of profile analysis, and also to analyze association of periodicity classes with ORFs that can be observed in DNA sequences belonging to these classes. In addition, if TP has a correlation with ORF, then we can reveal regions in genes where shift of ORF relative to TP has occurred or those regions where inversion of DNA bases has happened. Such capabilities are connected with a fact that periodicity cannot disappear because of individual evolution events such as deletions or insertions of nucleotides or inversion of DNA sequences (Korotkov et al., 2003a).

Building TPCs can be done well using method of information decomposition (Korotkov et al., 2003a,b) that allows introducing the term of TPC as 3×4-sized matrix. In a matrix, column attributes are positions in period, while row attributes are nucleotides (Korotkov et al., 2003a). Introduction of matrices simplifies joining similar TP into classes, searching of the shift between ORF and TP of some class, and also revealing of TP in inverted form.

In current work, we tried to solve three problems. First, we wanted to find all the coding sequences having TP in genes from KEGG databank (http://www.genome.ad.jp/kegg/) of version 29 (more than 4×10^5 genes) by information decomposition method and then to combine closely-related matrices of TP into classes. Therefore, for each gene under consideration we built corresponding matrix of TP that was bound to ORF of the given DNA region. This means that the first column in triplet periodicity matrix (TPM) corresponded to the first base of ORF presented in DNA region having TP.

Consequently, we were able to obtain 2520 TPCs by joining close matrices and taking into account possible cyclic shifts and inversions. These classes contain 94% of all TP regions found in genes. For each class we built the list of matrices which the given class consists of, and list of ORF positions. In this list, it was also indicated if the matrix was used in inverted form or not during joining the class. Each class in the list always contained one ORF that could be assumed as dominant for the class. This frame was common for majority of all TPMs joined by the class. We denoted this ORF as TPC–ORF. For each TPC we revealed genes having ORF different from class' one also including cases of inversion.

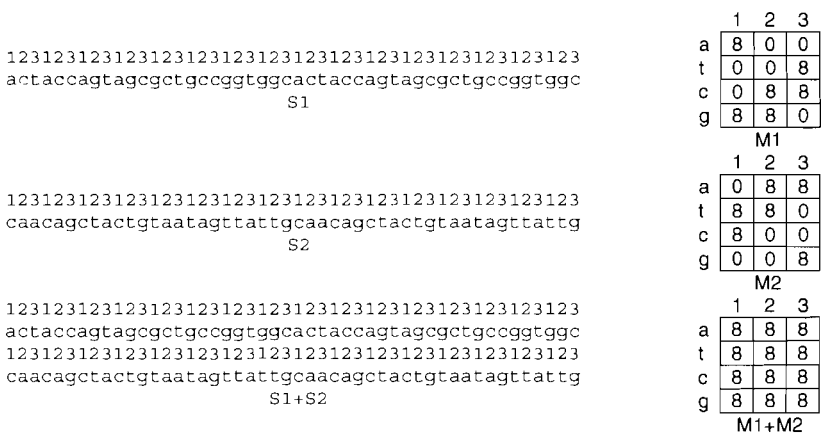

Fig. 2. Example of DNA base sequences with antagonistic TP. By antagonistic we assume periodicities disappearing after matrices join. Matrices M1 and M2 are built for TP in sequences S1 and S2. If we use sequences S1 and S2 in learning sample then their triplet periodicities will "annihilate" each other. Aggregation of matrices M1 and M2 illustrates this effect. After such join (matrix M3) a TP disappears.

Second, we wish to check whether hypothetical amino acid sequences translated from TPC–ORF have similarity with sequences from UniProt databank (http://www.uniprot.org/) or not. We checked those genes that had mismatch between their own ORF and the one of a TPC. This can be the evidence of the fact that ORF shift has occurred in the analyzed gene during its evolution. We confirmed existence of such a shift because we found similarity between hypothetical amino acid sequences and amino acid sequences from UniProt databank.

Third, we wanted to check the possibility of existence of amino acid sequences translated from inverted DNA regions. In order to do that, we selected those genes in which TPC was observed in inverted form only. We have built hypothetical amino acid sequences for these inverted gene sequences. In the current work we have found that many of these hypothetical amino acid sequences have high level of similarity with amino acid sequences from UniProt databank. This fact shows that shifts of ORF and inversions of DNA sequences are widespread in genes, and classes of TP can be useful tool for identification of such events.

2. Materials and methods

2.1. Search for triplet periodicity in genes

We performed a search for TP in genes accumulated in 29th release of KEGG databank by the method of information decomposition (Korotkov et al., 2003a,b). We used only coding sequences (CDS) for classification of triplet periodicity. These sequences started with start-codon and finished with stop-codon; introns have already been cut out. The total fraction of these sequences in the set of genes from KEGG-29 was equal to 97%, thus we have studied the most part of the genes. Namely, we compared DNA base sequence of each gene $A(n)=\{a(1)a(2),\ldots, a(n)\}$ with equal-sized artificial periodic sequence of the form: $S(3)=\{s(1)s(2)s(3)s(1)s(2)s(3),\ldots, s(1)s(2)s(3)\}$, where $s(1)\equiv$'1', $s(2)\equiv$'2', and $s(3)\equiv$'3'. In this sequence the symbols were treated as numbers. To compare these sequences, we filled coincidence matrix $M(3\times4)$. This matrix has symbols '1', '2' and '3' as attributes of columns and symbols of DNA base sequences $w(1)\equiv$'a', $w(2)\equiv$'t', $w(3)\equiv$'c', and $w(4)\equiv$'g' as attributes of rows. Element $m(i,j)$ of the matrix shows the number of $w(i)s(j)$ coincidences between two compared sequences. During filling matrix M, first base of first codon always corresponded to symbol $s(1)$ of artificial periodic sequence. After filling matrix M, we calculated mutual information according to formula (Kullback, 1978):

$$I = \sum_{i=1}^{3}\sum_{j=1}^{4} m(i,j)\ln m(i,j) - \sum_{1}^{3} x(i)\ln x(i) - \sum_{1}^{4} y(j)\ln y(j) + n\ln n \quad (1)$$

where n — length of examined symbolic sequence, $x(i)$, $i=1, 2, 3$ — frequencies of symbols '1', '2' and '3' in artificial periodic symbolic sequence; $y(j)$, $j=1, 2, 3, 4$ — frequencies of symbols in examined symbolic sequence. After calculation of mutual information, we estimated probability of random similarity between sequences $S(3)$ and $A(n)$. We used Monte-Carlo method to make such an estimation (Chaley et al., 1999) where statistics (Z value) was calculated as:

$$Z = (I-\bar{I})/\sqrt{D(I)} \quad (2)$$

where $\bar{I}$ and $D(I)$ are mean value and dispersion of mutual information value for a set of random matrices with same sums $x(i)$ and $y(j)$ as in source matrix M. Z value has a distribution close to normal. We checked this by comparing artificial periodic sequence with a set of random sequences with a length 10^7 symbols. This allows the use of Z as a measure of similarity between artificial periodic sequence and DNA base sequence. The higher Z value is, the higher is similarity between sequences S and A, and also periodicity is more explicitly expressed in sequence A. In addition, it is convenient to use matrix M for representing the form of periodicity observed in sequence A because same Z values can be obtained for different matrices M.

We scanned sequence A for a region having maximally expressed TP. We denote such a region as "maximal subsequence" of A. In order to find it, we examined all possible positions of left and right bounds for sequence $A(i)$, $n\geq i\geq 30$. We searched for sequence having maximal Z value. Sequences A and S were not changed or shifted relatively to each other; searching for maximal subsequence of A was accomplished only by shifting starting and ending coordinates of subsequence in A and by filling matrix M. We shifted left and right bounds by a step divisible by three bases. This means that first position of matrix M for maximal subsequence always corresponded to first base of codon and ORF. We assumed shift index for matrix M (Table 1) to be equal 1.

If Z value for maximal subsequence of A was greater than 5.0, then we assumed that we found region with TP. Value of Z greater than 5.0 ensures that probability of incidental revealing of TP in DNA base sequence is less than 10^{-6}. Thereafter, we saved maximal subsequence found for the given gene, its coordinates in the gene and periodicity matrix M reflecting a type of the found TP. Revealed sequences were saved (with auxiliary information) for further obtainment of TPCs.

We have chosen threshold level $Z>5.0$ for revealing TP in order to make number of incidentally found TP regions equal to about 1% of all regions with TP found in genes from KEGG-29 databank. To choose threshold Z value, we generated a set of random DNA sequences of the same size and with the same sequence length distribution as in genes from KEGG databank. For $Z>5.0$ number of found sequences was 7.200, i.e. about 1.5% of found regions with TP. For $Z>6.0$ we found 172 such sequences and none for $Z>7.0$. We have consciously chosen the level $Z>5.0$ in order to compose significant TPCs presenting in various genes as fully as possible. We checked stability of significant classes' classes' selection (see Section 2.2) by adding random matrices with TP in the range $0<Z<7.0$ into a set of found TPMs. Addition of up to 2% of random matrices did not change significant classes of TP.

We also checked if the triplet periodicity found in maximal subsequence of A was induced by periodicity with longer period divisible by three. To do this, we built the full specter of information decomposition (Korotkov et al., 2003a,b) for the maximal subsequence of A and revealed the period length for which the value of Z was maximal. If this length was not equal to 3, then the triplet periodicity of maximal subsequence of A was treated as induced one (Korotkov et al., 1997), thus it was excluded from further investigation.

Table 1
Transformations of TP matrix for various shift indices relatively to central class matrix

Matrix M after a shift	Shift index 1				Shift index 2				Shift index 3			
		1	2	3		**1**	2	3		**1**	2	3
	a	*m11*	*m21*	*m31*	a	*m31*	*m11*	*m21*	a	*m21*	*m31*	*m11*
	t	*m12*	*m22*	*m32*	t	*m32*	*m12*	*m22*	t	*m22*	*m32*	*m12*
	c	*m13*	*m23*	*m33*	c	*m33*	*m13*	*m23*	c	*m23*	*m33*	*m13*
	g	*m14*	*m24*	*m34*	g	*m34*	*m14*	*m24*	g	*m24*	*m34*	*m14*
	4				**5**				**6**			
		1	2	3		1	2	3		1	2	3
	a	*m32*	*m22*	*m12*	a	*m12*	*m32*	*m22*	a	*m22*	*m12*	*m32*
	t	*m31*	*m21*	*m11*	t	*m11*	*m31*	*m21*	t	*m21*	*m11*	*m31*
	c	*m34*	*m24*	*m14*	c	*m14*	*m34*	*m24*	c	*m24*	*m14*	*m34*
	g	*m33*	*m23*	*m13*	g	*m13*	*m33*	*m23*	g	*m23*	*m13*	*m33*

The table shows the positions of the elements of TPM M for six shift variants with which it can be merged into class. Shift index equal to 1 shows full equivalence of two matrices, i.e., absence of transformation in matrix M. This also means that ORFs in DNA sequences where matrix M were found and "central class matrices" (CCM) coincide with each other. Shift indices equal to 2 and 3 correspond to cyclic matrix shift after transformation per 1 and 2 bases, respectively. Shift index equal to 4 corresponds to inversion of matrix M relative to CCM. This means that in matrix M rows corresponding to *a* and *g*, *t* and *c* are swapped. Thereafter, columns 1 and 3 in matrix are also swapped. These transformations are designated by asterisk. Shift indices equal to 5 and 6 show correspondence of matrices analogous to the one with shift index equal to 4, but with cyclic column shift per 1 and 2 bases, respectively.

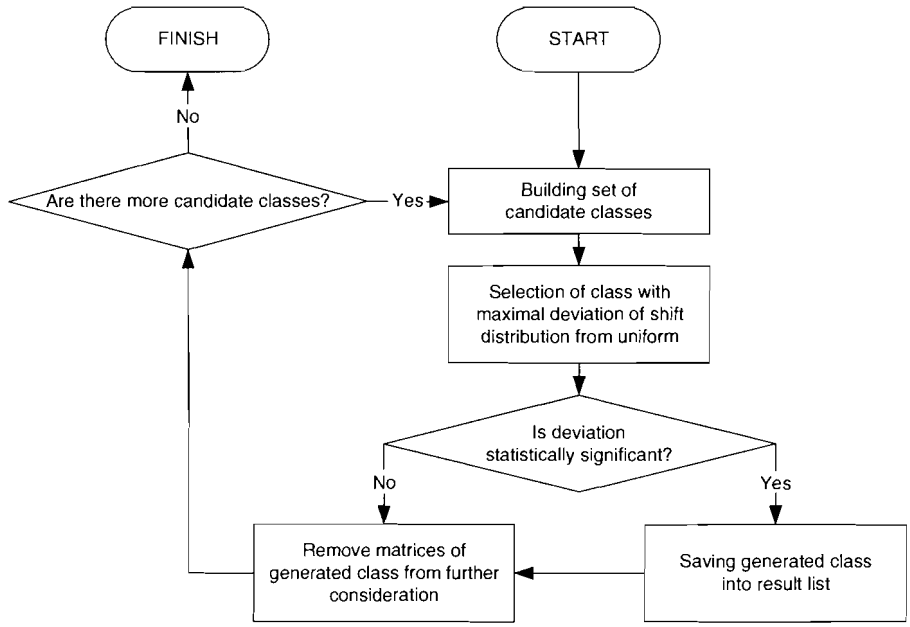

Fig. 3. Algorithm of classification.

In case of revealing TP region in a gene, we excluded it from the further consideration. If gene had fragments that were not included into the maximal subsequences, then they were processed again to search for more TP regions. Such an examination lets us find genes in which there are two or more TP sequences, including sequences with different matrices *M*.

2.2. Classification of triplet periodicity

We carried out classification of TPMs by sequential class forming. Classification process included a set of cyclically repeated steps (further referred as *iterations*). Every iteration forms one TPC. In turn, iteration was conducted in two *stages* described below.

During iteration, on the first stage we joined TPMs into candidate classes by using measure of their similarity to each other, allowing three direct cyclic shifts and three cyclic shifts in case of inversion. At the same time, for each candidate class we made a set of shift indices with which the matrix was joined into class. Thereafter, in each candidate class the most therein represented index was assigned to the value equal to 1, and other indices were cyclically shifted. For example, if the most represented shift index (relatively to central matrix) was equal to 2, then it was assigned a new shift index 1. In this case shift index 1 was replaced by 3, 3 by 2, 4 by 5, 5 by 6 and 6 by 4 (see Table 1). Consider also an example, where shift index 4 (inversion without cyclic shift) is dominant. Then we replace shift index 1 by 4, 2 by 5, 3 by 6, 4 by 1, 5 by 2, and 6 by 3. Consequently, the shift index which is the most represented in candidate class (relatively to central matrix) always has the value 1.

During second stage of class forming iteration, we chose the most nonrandom candidate class on the basis of uniformity degree of its shift index set. This means that candidate classes having approximately same numbers of each shift index (all possible are from 1 to 6, see Table 1) were recognized as random. Then matrices included in the most nonrandom candidate class were excluded from further examination and next iteration of forming a new class was performed. Scheme of class forming is shown in Fig. 3.

Let us consider each stage of iteration in detail. On the first stage of current iteration for each TPM we determined all matrices it is similar to. Each matrix forms candidate class consisting of itself and matrices similar to it. This matrix will be referred as CCM, and the set of the matrix will be referred as "candidate class". To estimate the degree of matrices' dissimilarity, the quantitative measure described below was introduced. During matrix comparison, all possible cyclic shifts and DNA sequence inversion were taken into account, i.e. totally 6 comparisons were made for a pair of matrices. At each comparison we fixed central class matrix and only the matrix it was compared with was transformed. From these 6 matrix pairs we chose the one with the smaller dissimilarity. For each candidate class a list of TPMs with corresponding shift indices (from 1 to 6) was built. Correspondence between shift indices and matrix transformation is shown in Table 1. We performed process of candidate class formation using each TPM as a central class matrix sequentially.

On the second stage of iteration, we chose candidate class having the greatest deviation of matrix shift distribution inside it from uniform one. All the matrices included in this class were excluded from further examination. Formation of the next class was performed for the remained remaining matrices. Classification process continued until candidate class containing at least two matrices (central class matrix and another one) could be formed. After classification, we

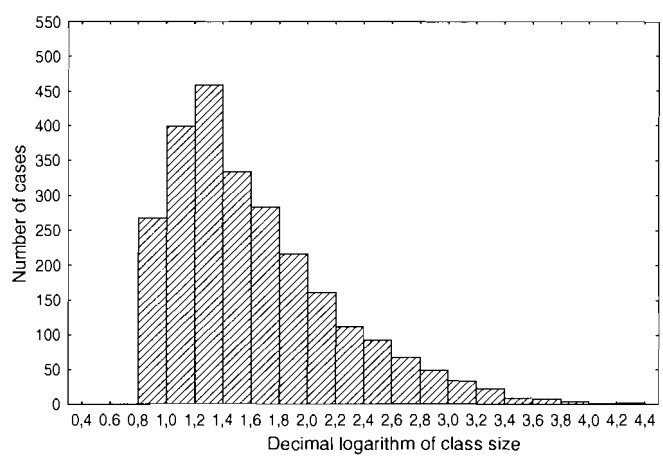

Fig. 4. Distribution of TPCs by size. Logarithmic axe *X* indicates class size and axe *Y* shows number of classes. We see that major number of class have size from 1 to 50 matrices.

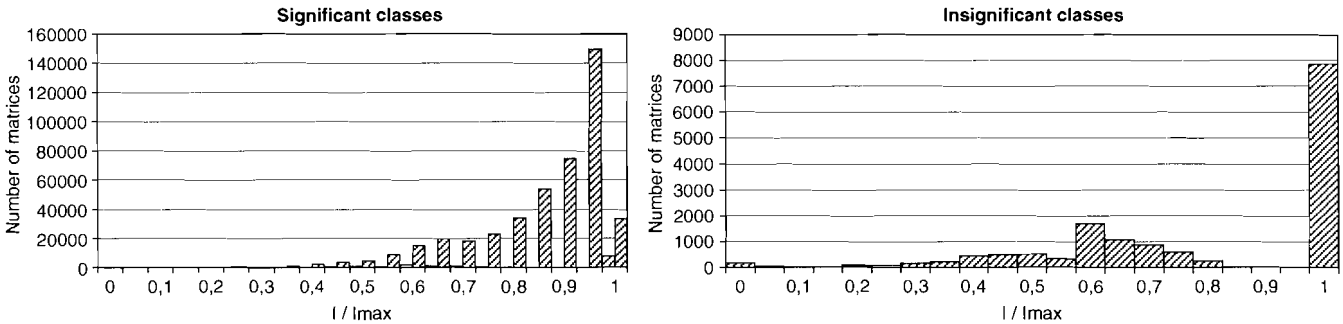

Fig. 5. Distribution of I/I_{max} value for significant and insignificant classes.

obtained the set of TPCs and set of matrices which were not included in any class. For each class a set of shift indices for TPMs included in this class was also formed.

Let us consider the process of two matrices' comparison and criterion of dissimilarity between them. As a criterion of dissimilarity between two matrices M^1 and M^2, we used the measure W (Gmurman, 2003) defined below as:

$$W = \sum_i \sum_j tij \tag{3}$$

A matrix $T=\{t_{ij}\}$ is defined as:

$$t_{ij} = \underline{m} \frac{\frac{m_{ij}^1}{y_j^1} - \frac{m_{ij}^2}{y_j^2}}{\sqrt{p(1-p)\left(\frac{1}{y_j^1} + \frac{1}{y_j^2}\right)}} \tag{4}$$

where $y_j^k = \sum_i m_{ij}^k$, $p=1/3$. Variable t_{ij} has approximately normal distribution. Variable W has χ^2 distribution with 8 degrees of freedom.

On forming candidate class, we introduced threshold value $W=W_0=3.44$. We consider that the matrix belongs to a candidate class if the value of W calculated by Eq. (3) is less than threshold value W_0. Value of W_0 asserts probability of accidental joining of matrices into one class to be less than 8.22×10^{-4}. It was obtained by Monte-Carlo method after generating random matrices and choosing the matrices with $Z>5.0$ among them (see Eq. (2)).

A choice of the value $W_0=3.44$ is related to two factors. First, we wanted to unite in classes as many matrices M as possible, and to leave outside the classes as smaller number of matrices as possible. At the same time, we wanted classes to be maximally representative and the number of classes to be relatively small. But, on the other hand, we wanted major number of matrices to be joined into nonrandom classes by homogeneity of shift indices presented in the class. The classification carried out for TPMs with values of W_0 higher than 3.44 showed that number of matrices included in nonrandom classes by shift index is decreasing under increasing W_0. From this point, the value $W_0=3.44$ appeared optimal because it provided sufficiently large number of TPMs that were included in nonrandom classes while number of formed classes remained relatively small.

On the second stage of classification iteration, we chose candidate classes that were maximally homogenous by shift index. We used informational criteria to do this. Let $x_1, x_2,\ldots, x_6$ be a number of shift indices 1, 2,..., 6 in class and N to be their total quantity, i.e. $\sum_{i=1}^6 x_i = N$.

By estimating shift indices homogeneity we check the hypothesis that a sample belongs to polynomial population $p_1,\ldots, p_6$, $\sum_{i=1}^6 p_i = 1$, where $p_i p_i = 1/6$. Discrimination information is calculated as (Kullback, 1978):

$$I = \sum_{i=1}^6 x_i \log \frac{x_i}{Np_i} = \sum_{i=1}^6 x_i \log x_i - \sum_{i=1}^6 x_i \log p_i - N \log N. \tag{5}$$

Variable $2I$ has a distribution χ^2 with 5 degrees of freedom (Kullback, 1978). Value of $2I$ equals zero if probabilities $f_i = \frac{xi}{N}$ are equal to the values p_i and $2I$ reaches maximum when value of f_i for some i is equal to 1.0 and other f_i values are equal to zero. This means that maximally heterogeneous shift index distribution in a class produces maximal value of $2I$. Consequently, we selected a class that had the maximum value of $2I$ among obtained classes, and then the process was repeated. Such a repetition continued until a class having value of $2I$ higher than some critical value $2I_0$ existed. Value of $2I_0$ was selected to obtain, on average, not more than one class as a result of making classification for each set of random matrices. This number is provided by choosing value of $2I_0$ equal to 40.5. We determined number of random classes that were obtained after classification of random matrices for selected threshold value of $2I_0$. Thereto we generated 30 sets of random TPMs having the same size as a set of matrices obtained under analysis of KEGG databank had. We have selected only the matrices having value of $Z>5.0$ (see Eq. (2)) for these sets. This means that approximately one of 10^6 generated random matrices was selected to be included in these sets. Then we conducted two stage classification described above for each of 30 sets. Results have shown that for these 30 sets of random matrices only 21 periodicity classes having $2I>2I_0$ can be built. So during such classification we can generate ~0.7 random classes per set on average. These random classes contain only 0.015% of all random matrices subjected to classification.

2.3. Analysis of heterogeneity of shift index distribution in insignificant classes

We can expect that not all of the classes built will have $2I$ $2I>40.5$. There are two reasons for such a case. First, class may have extremely heterogeneous shift index distribution, but value of N (Eq. (5)) may be small (few matrices in class). Second, shift index distribution (x_i in Eq. (5)) can be close to uniform. To separate these two cases, we must introduce a measure that would reflect degree of heterogeneity of

Table 2
Distribution of matrices' shift relative to TPC–ORF

Shift of periodicity against TPC–ORF	Number of matrices	Matrix shift indices
ORF 1	407687	1
ORF 2	2558	2
ORF 3	3162	3
Total direct with frame shift	5720	2+3
Total direct	413407	1+2+3
Inversion, ORF 1	20199	4
Inversion, ORF 2	7616	5
Inversion, ORF 3	2576	6
Total inverted with frame shift	10192	5+6
Total inverted	30391	4+5+6
Total with frame shift or inverted	36111	2+3+4+5+6
Total number of periodicity regions	443798	1+2+3+4+5+6

Table 3
Number of similarities for various ORF shifts

Variant of coding sequence transformation	Number of similarities			Shift index	Fraction in total number of translated periodic sequences
	By both ORFs	By class ORF only	By KEGG ORF only		
ORF 2	7	823	335	2	0.46
ORF 3	14	753	391	3	0.37
Inversion, ORF 1	25	441	12302	4	0.64
Inversion, ORF 2	12	352	3130	5	0.46
Inversion, ORF 3	0	38	840	6	0.34
Total	58	2407	16998		0.54

values x_i and would not depend on sample size N. For values $p_i p_i = 1/6$ Eq. (5) can be rewritten as:

$$I = \sum_{i=1}^{6} x_i \log 6x_i - N \log N. \quad (6)$$

In a case when x_i is also distributed uniformly ($x_i = N/6$) the value of $I = 0.0$. Maximum value of mutual information is observed for the case when some value of x_i is equal to N and remaining five values of x_i are equal to 0. In this case:

$$I_{\max} = N \log 6. \quad (7)$$

The ratio of the values

$$\frac{I}{I_{\max}} = 1 + \frac{\sum_{i=1}^{6} f_i \log f_i}{\log 6} \quad (8)$$

where $f_i = x_i/N$, will not depend on sample size N. Therefore, we can solve a problem of shift index distribution in insignificant classes by using value of $I/I_{\max}$ that does not depend on sample size.

3. Results

Totally we have analyzed 578868 genes accumulated in 29th release of KEGG databank (http://www.genome.ad.jp/kegg/). Overall number of TP regions was 472288. They were found in 457333 genes. These data show that more than 79% of genes have regions with TP. Such results conform to earlier works on revealing TP either by information methods or by other approaches (Tiwari et al., 1997; Azad and Borodovsky, 2004; Henderson et al., 1997; Grosse et al., 2000). For each DNA sequence with TP we calculated corresponding symbol coincidence matrix M. Then we combined these matrices into classes. We solve two tasks during classification. First, we can find the level of TPMs' diversity. Second, we can determine whether there exists linkage between first codon base and first TP position, and, if it exists, how stable it is. During process of finding TP by information decomposition method, all first positions of triplet matrix corresponded to first codon positions, and they could only be changed during joining matrices into classes when matrix cyclic shift or its inversion were possible. In order to study this linkage, we generated a set containing shift indices of all matrices included in TPC simultaneously with matrix of this class. Consequently, we obtained 2520 classes of periodicity matrices (http://victoria.biengi.ac.ru/ancorfs/classes.php). Classes have large size diversity, from 1 to tens of thousand (Fig. 4).

Classes with $2I > 40.5$ (see Section 2.2) contain 443798 of 472288 found TPMs. 8591 matrices were not included in any class and remained autonomous. 19899 matrices were included in insignificant classes (with $2I \leq 40.5$). We can see that about 94% of matrices were included in significant classes that show existence of linkage between TP and ORF. Class with $2I \leq 40.5$ could include small number of matrices, so that this number did not allow these classes to reach $2I > 40.5$, while shift index distribution for these classes could be heterogeneous. In order to check this hypothesis, we built distribution of $I/I_{\max}$ value for significant and insignificant classes (Fig. 5).

From this distribution we see that the value $I/I_{\max}$ for most classes lies in a range from 0.5 to 1.0. In a case of insignificant classes, they mostly lie in the same range, and about four thousand classes have $I/I_{\max}$ near or equal to unity. Therefore, in a case of insignificant classes a linkage between ORF and TPC is also observed. General conclusion of classification made is that there exists significant linkage between TPC and ORF in gene. We have also found that only about 8% of matrices were included in class with some shift or inversion of reading direction. This means that only 36111 of 443798 revealed periodicity cases in significant classes have ORF distinct from the common for majority of matrices in the class.

These data allow introducing of major ORF for each class — TPC–ORF (see Section 1). As it can be seen from the data shown above, such a frame for all classes is the ORF without shift or inversion (shift index for matrices equals 1). Then we examined total number of matrices included in significant classes with shift indices 2–6 (Table 2).

From Table 2 we can see that maximal number of matrices with ORF different from TPC–ORF have shift index equal to 4, i.e. they have inversion without cyclic shift. This corresponds to inversion of coding sequence of DNA bases with superposition of ORF without shift (Table 1).

```
224 KNPTNVKNVVKPSVFSVLFEGIKGLIVERNPMNVKNVEKPSFILQAFEHMXXYTLGLDLI 283
    + P  +  V +PSV       I   I+ERN  +VKNV K S   Q  + +  + +  +
 98 EKPYKLMIVARPSVICQPLHAIVDFILERNLTSVKNVMKLSVSNQVLKDIGEFIMERNCT 157

284 NVKNVGKPSTLLIPVECMKELILEKNHMNVNDVANHSVGPFLFDCMKELILERNLMSVNS 343
    NV +V +PS    P+  + + I+E+N  +V +V   +V   +   +++ ILERN++SV
158 NVMSVARPSVRSYPLPAIVDFIVERNLASVENVTRLTVSNKILKYVRKFILERNVISVMI 217

344 VIKPSVFQVPFENTKQLTLERNPMNVNNVVKPSVFQVPFKDMKGLTMQRNPMNVNSVGKP 403
    V + SV + P        +ERN  NV NV+K SV     KD+   T+ RN   V  V +P
218 VARSSVIRHPLYTIINFIVERNLTNVKNVMKLSVSNQTLKDIGEFTLVRNLTGVIGVARP 277

404 SGVQVIFEFPMKGHTLERNPMNVNSVEKFSFVPVPFFDCMKEHTLERNPMNVNYAVKPSVFQ 463
    S +      +    LERN  NV +V K S        + E  +ERN  +V    +PSV
278 SVIHHPLHAIIDFILERNLTNVKNVMKLSDTNQILKDIGEFIMERNRTSVMSVARPSVRS 337

    VPFENMKKFTLEISLLSVSNVVRPS 488
         +  F LEI+L+SV +V RPS
    HALHAIIDFILEINLISVMSVARPS 362
```

Fig. 6. Similarity of hypothetic amino acid sequence obtained by 2nd ORF (F2) from to locus *ZN525_HUMAN* from UniProt databank.

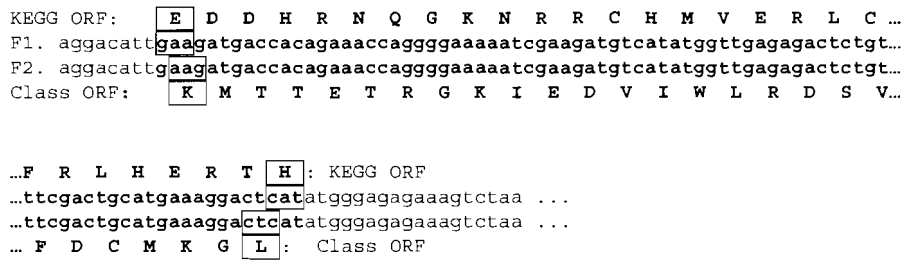

Fig. 7. Recoding DNA base sequence of locus *7561* from KEGG databank from 157th to 1911th nucleotide (F1) and from 158th to 1909th nucleotide (F2) into amino acid sequence. Figure shows beginning and end of sequences only.

On the second place according to the number of matrices we can see the case of matrix inversion with superposition of codons with shift by one base. Then we observe similarity of matrices with cyclic shift only (shift indices 3 and 2) and, finally, we have similarity of matrices with inversion and shift by 2 bases.

We have also checked periodicity matrices for existence of symmetry inside them. This could be a cause of found shifts in classes. In order to check this hypothesis, we compared each matrix to itself after conducting all possible cyclic shifts and inversion. We found that 0.4% of class matrices have similarity to itself (to one of five its own variants obtained by shift and inversion). Similarity estimation was conducted by the same criteria as we used during classification. It looks like that though symmetry occurs in some matrices, in general it does not play significant role in forming shifts between matrices inside classes (0.4% of symmetric matrices against 8% of matrices with shift).

Data of Table 2 show that ORF shifts and DNA sequence inversions occur in DNA coding sequences. It is difficult to change the TP by individual mutations (Korotkov et al., 2003a). Therefore, we can observe it as a trace of previously existed ORF changed by deletions, insertions or inversion of DNA sequences in genes. However, this hypothesis needs additional proof. We can prove it by recoding nucleic sequence, where we found ORF shift, into amino acid sequence by using TPC-ORF. Same translation can be done for sequences where TPM with shift indices 4, 5 and 6 was obtained. We must also add inversion procedure for DNA sequence, i.e., its flip-over by 180 degrees and replacement of DNA sequence bases to complementary ones. After such transformation we obtain hypothetic amino acid sequence that could be in gene before ORF shift or sequence inversion took place. If we show that hypothetic amino acid sequence has homologous sequences in UniProt databank, then this will prove that processes of ORF shifts and inversions actually took place in DNA fragment. At the same time, this fact can be considered as very probable if amino acid sequence obtained by original ORF from KEGG databank (KEGG ORF) also has amino acid similarity.

To check this hypothesis, we searched for homologues of protein products coded in found gene regions. We performed coding by either KEGG ORF or class frame. Search for homologues in UniProt databank was conducted by using BLAST program. Only significant cases having probability of incidental coincidence (*e*-value) less than 5% were considered among found similarities. Then we chose the similarities for the amino acid sequences in which m_repeats program had not found amino acid repeats (Pellegrini et al., 1999). This allowed excluding the similarities which might have originated not due to common evolutional origin of these sequences, but due to presence of amino acid repeats in them. Each examined DNA sequence having shift index from 2 to 6 was recoded into pair of amino acid sequences according to existent and hypothetical (TPC) ORFs. Consequently, after conducting search we obtained lists of homologues for 20733 DNA regions with TP. For 16648 periodicity regions no significant similarity to their protein products obtained by any ORF was found, and also no repeats in similar sequences were found. In 58 cases we observed similarity for both hypothetic and actual amino acid sequences (Table 3).

Table 4
Matrix of TPC containing TP of DNA base sequence from 157th to 1911th nucleotide of locus ***7561*** from KEGG databank

	1	2	3
A	97	91	86
T	50	116	77
C	62	73	86
G	123	52	83

Example of found similarity for hypothetic amino acid sequence obtained by TPC-ORF is shown in Fig. 6 and is available in Internet at http://victoria.biengi.ac.ru/ancorfs/perinfo.php?perid=355386.

Such similarity was found for the gene with identifier *7561* from KEGG databank and the amino acid sequence from UniProt databank with identifier *Q8N782* (Ota et al., 2004). Gene *7561* has length of 1929 nucleotides; TP is most expressed in sequence from 157 to 1911 nucleotides. TPM of this DNA region was included in the corresponding class with shift index equal to 2 (Fig. 7).

It corresponds to 2nd ORF (first ORF corresponds to original, i.e. KEGG, amino acid coding in this gene). The TPC (Table 4) joined 2246 TPMs, and in the given class totally 96.7% (2172 of 2246) of matrices have shift index equal to 1.

Using this gene with ORFs 1 and 2 (Fig. 7) we built amino acid sequences. Then we analyzed similarity between these amino acid sequences and UniProt databank by BLAST program. We found 31 cases of similarity for amino acid sequence coded by TPC-ORF (hypothetic amino acid sequence) and 1427 cases of similarity for a protein coded by KEGG ORF. Best similarity with original (KEGG) ORF was found in the sequence *P17017* (Thiesen, 1990) (similarity from 53 to 637 amino acids), where full similarity (100%) between amino acid sequences was observed (Fig. 6).

Thus, data from Table 3 show that we can simultaneously reveal similarity for both class and KEGG ORFs only for few genes for which TP was joined into TPC with ORF shift. Probably, genes have accumulated many substitutions after an ORF shift took place, and similarity cannot be revealed or this sequence has no similar sequences in UniProt at all. Nevertheless, 2489 genes have similarity for amino acid sequences coded by TPC-ORF, but have no similarities for amino acid sequences created by KEGG ORF. These data suggest that at least 2489 sequences could be formed by ORF shifts or inversions.

We have developed databank that contains information on found similarities for all examined periodicity regions (http://victoria.biengi.ac.ru/ancorfs/), class matrices and genes included in TPCs (http://victoria.biengi.ac.ru/ancorfs/classes.php).

4. Discussion

Phenomenon of TP was observed long ago (Fickett, 1998; Staden, 1994; Baxevanis, 2001; Gutiérrez et al., 1994; Gao et al., 2005; Yin and Yau, 2007; Eskesen et al., 2004; Bibb et al., 1984; Konopka, 1994; Trifonov, 1999). The data we obtained shows that major fraction (94%)

of TP found in genes from KEGG-29 databank can be reduced to 2.520 classes. Each class on the average contains 86.2% of matrices with first base of matrix *M* corresponded to first base of codon, and only few classes (about 2%, 53 of 2520) contain less than 50% of matrices with shift index equal to 1. Therefore, we can affirm that there exists strong correlation between TP and ORF in gene. If such correlation does not exist, then formation of TPCs by the algorithm used will be impossible. We showed this during classification of 30 random sets of TPMs obtained for $Z > 5.0$.

Introduction of TPCs seems to be important in development of more effective algorithms for finding DNA coding regions. During classification we characterized TP in detail and antagonistic matrices were included into different classes and do not "kill" each other. Under "antagonistic matrices" we assume such matrices that after aggregation (first column is joined with first one, second with second one, and so on) they give degree of TP less than the one of source matrices (Fig. 2). This allows HMM methods and profile analysis to reveal coding sequences at more significant level based on TP belonging to antagonistic classes. The linkage between TP and ORF allows us not only to make predictions whether new nucleotide sequences belong to coding regions or not, but also to determine the most possible ORF for them.

There are many classes having TPMs shifted or inverted relatively to original ORF of a gene. This fact witnesses for two hypotheses. First, such shift or inversion could be a consequence of nucleotide deletions and insertions or DNA sequence inversions when new ORF and new amino acid sequence arose (Raes and Van de Peer, 2005; Hahn and Lee, 2005). Since latent TP appears as collective property, it cannot disappear due to individual insertions, deletions, inversions or nucleotide substitutions (Korotkov et al., 2003a). Thus in gene with nucleotide deletion and insertion or with nucleotide sequence inversion in underlying sequence, a TP linked to ancient ORF will exist. A new ORF will be formed for this sequence. After detection of TP in such nucleotide sequence and its further classification, it can be included in corresponding class with shift or inversion relative to new ORF. We actually detected these facts in the current work. This hypothesis is partially confirmed by data of Table 3 and by presence of homologues to sequences built using ORF of corresponding class in UniProt databank.

Secondly, we cannot eliminate the fact that TP of a same type can be incidentally formed in different genes and in different ORFs. In this case, during classification such TPMs will be included in corresponding class with different shifts or inversion. This can result in revealing some matrices in TPC which joined the class with a shift. We cannot easily estimate the fraction of such matrices in classes, but probably it is not too high because there are many TPCs which do not contain at all or contain just some matrices with a shift.

We did not exclude triplet periodicity of genes which had similar nucleotide sequences from classification process. When using our classification algorithm, this can be the cause of size inflation for some triplet periodicity classes due to presence of homologous sequences. However, it is better not to exclude the similar sequences because the similarity level can be either high or low. On the one hand, choosing the similarity threshold level is rather arbitrary process that can hardly be sensible, but, on the other hand, this value would have a great influence on class formation. We think that exclusion of similar sequences can only insignificantly decrease the number of classes, but cannot substantially change the results obtained. In any case, the classification performed by us includes the whole variety of triplet periodicity which is present in the genes from KEGG-29 database. In this sense, the classification is exhaustive. Moreover, the increased representation of triplet periodicity from similar sequences is not very important for the goals of our investigation since it cannot prevent us from using the obtained class matrices for more precise identification of DNA coding regions and for revealing the possible reading frame shifts in genes.

It is interesting to consider the possible origins of triplet periodicity in genes. The factors that can lead to formation of such a periodicity include the protein saturation with certain amino acids, use of synonymic codons (Tiwari et al., 1997), and also the certain order of triplet interchange (Sánchez and López-Villaseñor, 2006). It is also supposed that triplet periodicity can reflect the evolutional process of coding sequences' origination (Eigen and Winkler-Oswatitsch, 1981; Eskesen et al., 2004). The periodicity classes obtained by us integrally reflect all these characteristics of coding sequences. It can be also surmised that triplet periodicity of coding sequences causes the formation of certain specter of DNA natural vibrations (Bour et al., 2005; Girirajan et al., 1989; Chou, 1984). It is likely that these vibrations can play a certain role for performing some DNA–protein interactions.

References

Antezana, M.A., Kreitman, M., 1999. The nonrandom location of synonymous codons suggests that ORF-independent forces have patterned codon preferences. J. Mol. Evol. 49 (1), 36–43.

Azad, R.K., Borodovsky, M., 2004. Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory. Brief. Bioinform. 5 (2), 118–130.

Baxevanis, A.D., 2001. Predictive methods using DNA sequences. Methods Biochem. Anal. 43, 233–252.

Bibb, M.J., Findlay, P.R., Johnson, M.W., 1984. The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. Gene 30 (1–3), 157–166.

Bour, P., Andrushchenko, V., Kabelác, M., Maharaj, V., Wieser, H., 2005. Simulations of structure and vibrational spectra of deoxyoctanucleotides. J. Phys. Chem. B 109 (43), 20579–20587.

Chaley, M.B., Korotkov, E.V., Skryabin, K.G., 1999. Method revealing latent periodicity of the nucleotide sequences modified for a case of small samples. DNA Res. 6 (3), 153–163.

Chou, K.C., 1984. Low-frequency vibrations of DNA molecules. Biochem. J. 221 (1), 27–31.

Eigen, M., Winkler-Oswatitsch, R., 1981. Transfer-RNA: the early adaptor. Naturwissenschaften 68, 217–228.

Eskesen, S.T., Eskesen, F.N., Kinghorn, B., Ruvinsky, A., 2004. Periodicity of DNA in exons. BMC Mol. Biol. 5, 12.

Fickett, J.W., 1996. The gene identification problem: an overview for developers. Comput. Chem. 20 (1), 103–118.

Fickett, J.W., 1998. Predictive methods using nucleotide sequences. Methods Biochem. Anal. 39, 231–245.

Gao, J., Qi, Y., Cao, Y., Tung, W.W., 2005. Protein coding sequence identification by simultaneously characterizing the periodic and random features of DNA sequences. J. Biomed. Biotechnol. 2, 139–146.

Girirajan, K.S., Young, L., Prohofsky, E.W., 1989. Vibrational free energy, entropy, and temperature factors of DNA calculated by a helix lattice approach. Biopolymers 28 (11), 1841–1860.

Gmurman, V.E., 2003. Teoriya veroyatnostej i matematicheskaya statistika. Vysshaya shkola, Moscow (in Russian).

Gribskov, M., Veretnik, S., 1996. Identification of sequence pattern with profile analysis. Methods Enzymol. 266, 198–212.

Grosse, I., Buldyrev, S.V., Stanley, H.E., Holste, D., Herzel, H., 2000. Pacific Symposium on Biocomputing. Abstract Book, Hawaii, USA.

Gutiérrez, G., Oliver, J.L., Marín, A., 1994. On the origin of the periodicity of three in protein coding DNA sequences. J. Theor. Biol. 167 (4), 413–414.

Hahn, Y., Lee, B., 2005. Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. Bioinformatics 21, 186–194.

Henderson, J., Salzberg, S., Fasman, K.H., 1997. Finding genes in DNA with a Hidden Markov Model. J. Comput. Biol. 4, 127–141.

Issac, B., Singh, H., Kaur, H., Raghava, G.P.S., 2002. Locating probable genes using Fourier transform approach. Bioinformatics 18 (1), 196–197.

Karlin, S., Bucher, P., 1992. Correlation analysis of amino acid usage in protein classes. Proc. Natl. Acad. Sci. U. S. A. 89 (24), 12165–12169.

Konopka, A.K., 1994. Sequences and codes: fundamentals of biomolecular cryptography. In: Smith, D. (Ed.), Biocomputing: Informatics and Genome Projects. Academic Press., San Diego:, pp. 119–174.

Korotkov, E.V., Korotkova, M.A., Tulko, J.S., 1997. Latent sequence periodicity of some oncogenes and DNA-binding protein genes. Comput. Appl. Biosci. 13, 37–44.

Korotkov, E.V., Korotkova, M.A., Frenkel, F.E., Kudryashov, N.A., 2003a. The informational concept of searching for periodicity in symbol sequences. Mol. Biol. (Mosk.) 37 (3), 436–451.

Korotkov, E.V., Korotkova, M.A., Kudryashov, N.A., 2003b. Information decomposition method for analysis of symbolical sequences. Phys. Lett. A 312 (3–4), 198–310.

Kullback, S., 1978. Information Theory and Statistics. Peter Smith, Gloucester.

Makeev, V.Ju., Tumanyan, V.G., 1996. Search of periodicities in primary structure of biopolymers: a general Fourier approach. Comput. Appl. Biosci. 12 (1), 49–54.

Ota, T., et al., 2004. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat. Genet. 36, 40–45.

Pellegrini, M., Marcotte, E.M., Yeates, T.O., 1999. A fast algorithm for genome-wide analysis of proteins with repeated sequences. Proteins 35 (4), 440–446.

Raes, J., Van de Peer, Y., 2005. Functional divergence of proteins through frameshift mutations. Trends Genet. 21, 428–431.
Sánchez, J., López-Villaseñor, I., 2006. A simple model to explain three-base periodicity in coding DNA. FEBS Lett. 580, 6413–6422.
Snyder, E.E., Stormo, G.D., 1993. Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. Nucleic Acids Res. 21, 607–613.
Staden, R., 1994. Staden: statistical and structural analysis of nucleotide sequences. Methods Mol. Biol. 25, 69–77.
Thiesen, H.J., 1990. Multiple genes encoding zinc finger domains are expressed in human T cells. New Biol. 2 (4), 363–374.
Thomas, A., Skolnick, M.H., 1994. A probabilistic model for detecting coding regions in DNA sequences. IMA J. Math. Appl. Med. Biol. 11 (3), 149–160.
Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., Ramaswamy, R., 1997. Prediction of probable genes by Fourier analysis of genomic sequences. Comput. Appl. Biosci. 13 (3), 263–270.
Trifonov, E.N., 1987. Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences. J. Mol. Biol. 194, 643–652.
Trifonov, E.N., 1999. Elucidating sequence codes: three codes for evolution. Ann. N. Y. Acad. Sci. 870, 330–338.
Yin, C., Yau, S.S., 2007. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J. Theor. Biol. 247, 687–694.
Zhang, J., 2005. On the evolution of codon volatility. Genetics 169 (1), 495–501.
Zoltowski, M., 2007. Is DNA code periodicity only due to CUF — Codons Usage Frequency? Conf. Proc. IEEE Eng. Med. Biol. Soc. 1, 1383–1386.